# Sufficient Encoding of Dynamical Systems

From the grasshopper auditory system to general principles

DISSERTATION

zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat.)
im Fach Biophysik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät I
Humboldt-Universität zu Berlin

von
Herr Felix Creutzig
geboren am 9.11.1979 in Hannover

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Dr. h.c. Christoph Markschies

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:
Prof. Dr. Christian Limberg

Gutachter:

1. Prof. Dr. Andreas V. M. Herz
2. Prof. Dr. Naftali Tishby
3. Prof. Dr. Laurenz Wiskott

| | |
|---|---|
| eingereicht am: | 3. September 2007 |
| Tag der mündlichen Prüfung: | 14. Januar 2008 |

# Widmung

Meinen Eltern

# Contents

# Chapter 1

# Introduction

When you, as a member of the audience, listen to an early Beethoven sonata, you will automatically have a feeling for what accord or even motif will come next. A later Beethoven string quartet, however, will contain more surprising elements and you will not necessarily have a fixed expectation of the upcoming motif. Far more, this is true for twelve-tone music of Arnold Schönberg that leaves the listener with uncertainty. In fact, this unpredictability makes it difficult for the untrained ear to deal with this music, while the very same property creates a challenge for music enthusiasts.

As neuroscientists, we naturally ask for the neural basis of this phenomenon. Supporters of the efficient coding hypothesis state that neural systems are designed such that redundancy is reduced and the neurons' output is independent, conditioned on the input. This perspective is opined by Attneave, Barlow, Laughlin and Olshausen beside others. Specifically this implies that only those signal components are transmitted that cannot be predicted by other signal components that are simultaneously – or were previously – transmitted. Hence one can utilize available information to predict incoming signals and encode only those aspects of the incoming signal that were unexpected. From this perspective, efficient coding can also be called predictive coding. Mostly, neuroscientists have applied these ideas on spatial prediction in the visual system (**??**). In the auditory system, certain psychoacoustic observations can best be grasped by assuming a specific kind of predictive coding (**?**).

There is a second aspect of coding predictive information: Prediction may be required by the behaving organism. Consider the goalkeeper at a penalty shoot-out. The football may not need more than 300 ms to reach the goal. Hence the goalkeeper has a decisive advantage if he successfully predicts the correct corner by observing the movement of the football player approaching the penalty spot. Rather than an exception, restricted to high

performance sports, this kind of prediction is a common property of behavioral interactions between organisms. You will probably be acquainted with the situation where you try to concentrate on your work but are perpetually distracted by the tedious fly revolving around your head. If you, as an experienced fly catcher, want to kill the fly, you will not try to slam the animal at its current position but where it will be at the moment your hands meet. In other words, the art of fly catching is based on correctly predicting the fly trajectory. Even more, an evolutionary point of view suggests that organisms are only interested in information that can influence future action. Hence extracting predictive information may actually be not only a nice add-on but a cornerstone of sensory processing. Such a point of view is advanced by theoretical neuroscientists, e.g., Naftali Tishby and William Bialek, and experimental neuroscientist, e.g., Rodolfo Llinas, alike.

Both perspectives on predictive coding can be seen as complementary. However, they lead to distinct kind of questions. The efficient coding perspective emphasizes the question of data compression. The behavioral perspective, moreover, asks for extraction of the most predictive components of the incoming signal. Crucially, this allows the interpretation that not all information that is predictive necessarily needs to be encoded. Rather one can postulate that only information that is needed to perform a task is extracted. For example, when clapping your hands in order to grasp the fly, you need to estimate the approximate future location of the fly up to an order of magnitude of the size of your hands but not more. Indeed, why should an organism encode more information than can be used for motor action? At the best, this is a waste of ressources, at the worst it distracts from essential action. I suggest that an appropriate term for this additional facet is sufficiency – or for our purposes sufficient coding.

Does this mean that the notion of optimality becomes negligible? Of course, not. It will become clear that sufficient coding can mathematically be treated as a two-dimensional optimality problem leading to an optimal curve instead to a single optimum. In one dimension, one tries to maximize the accuracy of the representation, in the other dimension one tries to minimize the complexity of the model or coding costs of the system. In fact, ultimate perfection in one dimension may actually mean complete collapse in another dimension[1]. To emphasize this argument, I consider it necessary to use a therewith concordant terminology, i.e., sufficiency.

In this work, we will use two approaches to study coding and processing of

[1] An interesting illustration of this relation can be observed in economics. An exclusive focus on maximization of economic throughput as measured by economic growth in a ressource and sink limited environment leads to an overuse of natural assets with negative consequences for overall affluence.

temporal patterns – or equivalently – dynamical systems. First, we will focus on a particular sensory system, the auditory system of the grasshopper. We will analyze the processing of behaviourally relevant communication signals in a small neural network. In particular, we will gain insight how some relevant information about the signal, i.e., the ratio between alternating syllable and pauses, can be identified while getting rid of unwanted information such as the overall time-scale of the signal. This invariance computation can be viewed as a particular instance of sufficient coding in a setting where sensory processing and behavioural output is tightly coupled.

Inspired by the study of this exemplary neural system, we try to find a mathematical framework for information processing of temporal patterns. Technically, we seek to find a variable that maximizes the information that the past carries about the future while keeping the information rate low. The problem requires the information-theoretic treatment of the theory of dynamical systems. Effectively, the problem of efficient predictive coding can be mapped onto a particular instance of system identification belonging to the so-called subspace-based methods. Furthermore, the problem of finding a sufficient system in the sense that only the most predictive components are encoded can be identified with model reduction of dynamical systems.

In the following, we will provide a guideline of what to expect in the individual chapters of this thesis.

In chapter 2, we will introduce the auditory system of the grasshopper and investigate the spike train of one specific interneuron in response to natural occuring and artificially modulated mating signal. We will show that this neuron can encode one particular temporal feature of the communication signal, pause duration, by intraburst spike count. We will discuss this result in the context of burst coding in sensory systems.

In chapter 3, we postulate a putative mechanism that can read out this bursting neuron in a time-scale invariant manner. This is a desirable property for poikilothermic grasshoppers as their communication signal scale with outside temperature. Indeed, behavioral response is rather dependent on syllable to pause ratio but not on absolute syllable or pause duration.

In chapter 4, we model a minimal circuit simulating the spike train response of the bursting neuron. The main feature of this circuit is an interplay between fast excitation and slow inhibition. We show that such a model can also explain the response of neurons in the auditory forebrain of songbirds to vocal communication signals. We discuss the general properties of this ubiquitous circuit in auditory systems.

In chapter 5, we suggest an extended model of the grasshopper's auditory system that can detect communication signals comparably to results from behavioral experiments. The bursting neuron is an integral part of this larger

circuit. We show how the validity of this model could be tested in behavioral experiments.

In chapter 6, we introduce some basic results from information theory in order to put subsequent results into a broader perspective. From a neuroscientific point of view it is important that source and channel coding, i.e., data compression and data transmission, can be treated within one framework, similarly to information processing in sensory systems. Furthermore rate-distortion theory provides a first insight into the tradeoff between two contradicting information-theoretic objectives. With this background, we introduce the information bottleneck method, the method of choice from hereon.

In chapter 7, we define predictive coding for a discrete-time stochastic process with a Markov property, i.e., where – given the current state – previous states are irrelevant for predicting future states. We show that an information-theoretic approach provides an algorithm for extracting predictive components of the signal that is equivalent to linear slow feature analysis, another method that can model receptive field properties along the visual system. This result is important as it consolidates the idea of predictive coding by relating predictive coding to other established methods.

In chapter 8, we introduce the theory of dynamical systems. First, important terminology is clarified. Second, an overview over system identification methods with emphasis on subspace-based identification is given. Third, a short primer on model reduction is provided. Doing so, we obtain the groundwork and context to understand the scope of the subsequent results.

In chapter 9, we extend the approach of chapter 7 to all time steps, by this motivating the role of the state space as the information bottleneck between past and future of dynamical systems. We also obtain a particular variant of subspace-based system identification. In the main section, we directly apply the information bottleneck ansatz to linear dynamical systems, denoting this as the past-future information bottleneck. We derive necessary conditions for the so-called Hankel singular values such that the reduced system lies on the optimal information curve trading model accuracy against model complexity. We demonstrate the feasibility of the resulting algorithm chosing a spring-mass system as an example.

In chapter 10, we jointly discuss the results from the grasshopper auditory system and the past-future information bottleneck. We use an information bottleneck algorithm to extract predictive features from grasshopper communication signals and compare them with response properties of auditory neurons. The outlook illustrates the limitations of past-future information bottleneck and suggests future directions of research.

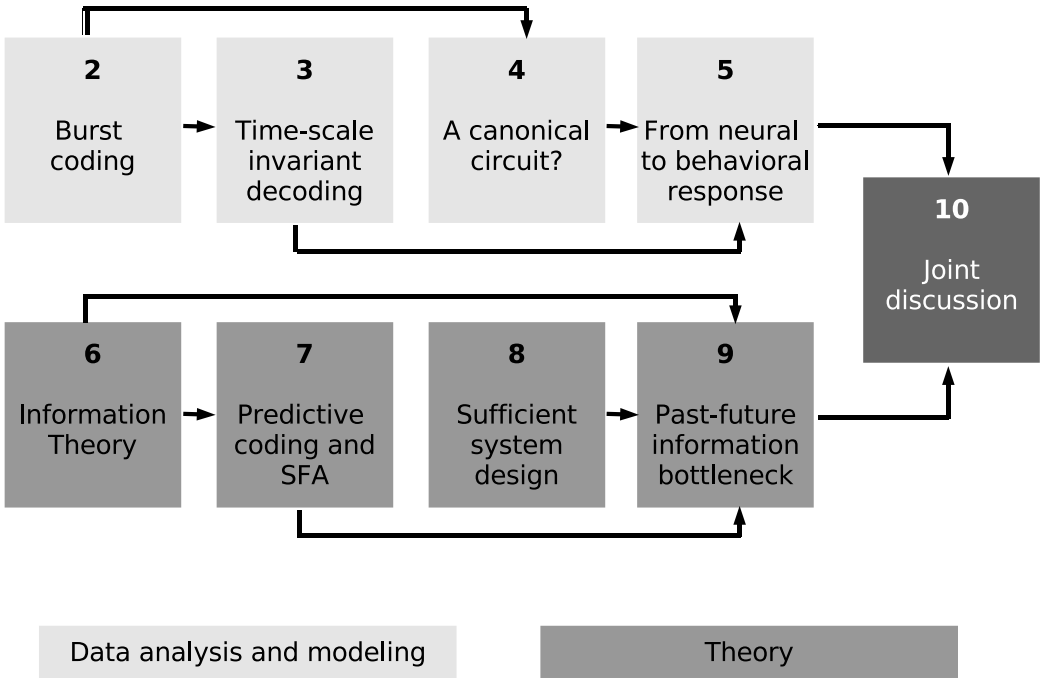


Figure 1.1: **A short guide how to read this thesis.**

# Chapter 2

# Burst encoding of a communication signal

Encoding the natural environment is the task of sensory systems. The potential data load of the environment is huge, and the organism must employ specific strategies to make sense of the incoming signals. These strategies are implemented into a set of neural coding schemes. The basic unit of the neural code is the spike, i.e., an action potential that may transmit information to other neurons. In sensory systems, spike trains, sequences of action potentials, encode and represent the external world (**?**). Several coding schemes have been suggested. In rate coding, information is carried with the firing rate, i.e., the number of spikes per time (**?**). Rate coding has enjoyed a prominent role in neurosciences for many decades. In agreement with the rate coding perspective, it has been shown that neurons respond to the summed activity of many synaptic inputs and act as integrators (**??**). However, there is evidence that speed of sensory processing limits the time available to read out spike trains (**?**), limiting the feasibility of rate codes (**?**). Hence, other scientists suggest that neurons have to be thought as temporal coincidence detectors, emphasizing the need for precise timing in the neural code (**??**). Additionally, information may be carried in the order of incoming single spikes in populations of neurons (**??**). All these suggestions and results require population codes. However, in some steps of sensory processing hierarchies the information flow converges onto a smaller number of neurons, constituting a neural information bottleneck. Hence, a small number of neurons must adapt to the challenge of encoding the behaviourally relevant features of the input signals. But how can a single neuron rapidly transmit information on quantitative properties of an external stimulus?

Here, we suggest that bursts, i.e., a series of action potentials within a short time scale, are ideally suited to rapidly transmit information in a

quantitative manner. The attractive property of bursts is that they can use two different codes simultaneously: the identity of a particular event by the burst's being and quantitative features by burst duration or intraburst spike count.

In this chapter, we will study a particular burst code. The system of our choice is the grasshopper auditory system. As particular attractive features, this system A) is sufficiently simple such that individual neurons can be analyzed and B) has neural responses that can be related to the animal's behaviour. Furthermore, C) the system is also complex enough such that interesting computational strategies can be observed.

In detail, we will focus on the following questions. Can bursts be used to classify temporal signals? What is the average signal preceding bursts with given spike count? What (behaviorally relevant) temporal signal feature is encoded in bursts? How much information is transmitted by each burst about the temporal signal? However, as we will rely on the auditory system of the grasshopper also for subsequent chapters, we first introduce anatomical and physiological characteristics of the auditory pathway and properties of the behaving animal. Later on, we provide an extended discussion of bursts in sensory systems.

## 2.1 Grasshopper fundamentals

### Anatomy and physiology

The anatomy of the auditory system (Fig. **??**) constrains the processing of sensory inputs. A tympanic membrane is located on each side of the lateral abdomen. About 70 spiking receptor cells are attached to each membrane (**?**). Four different kinds of receptor cells can be distinguished from their different response characteristics. Three of these receptor types are most sensitive to low carrier frequencies, the fourth responds most strongly to high carrier frequencies (**??**). The transduction from acoustic signals to receptor response has been extensively analyzed (**?????**). As long as a signal contains frequencies in the appropriate range, its amplitude distribution is well encoded by receptor neurons (**?**). The receptor cells project into the metathoracic ganglion where information is preprocessed before being sent into the head ganglion. As the highest neural processing stage, the head ganglion integrates available information and gives rise to behavioural responses.

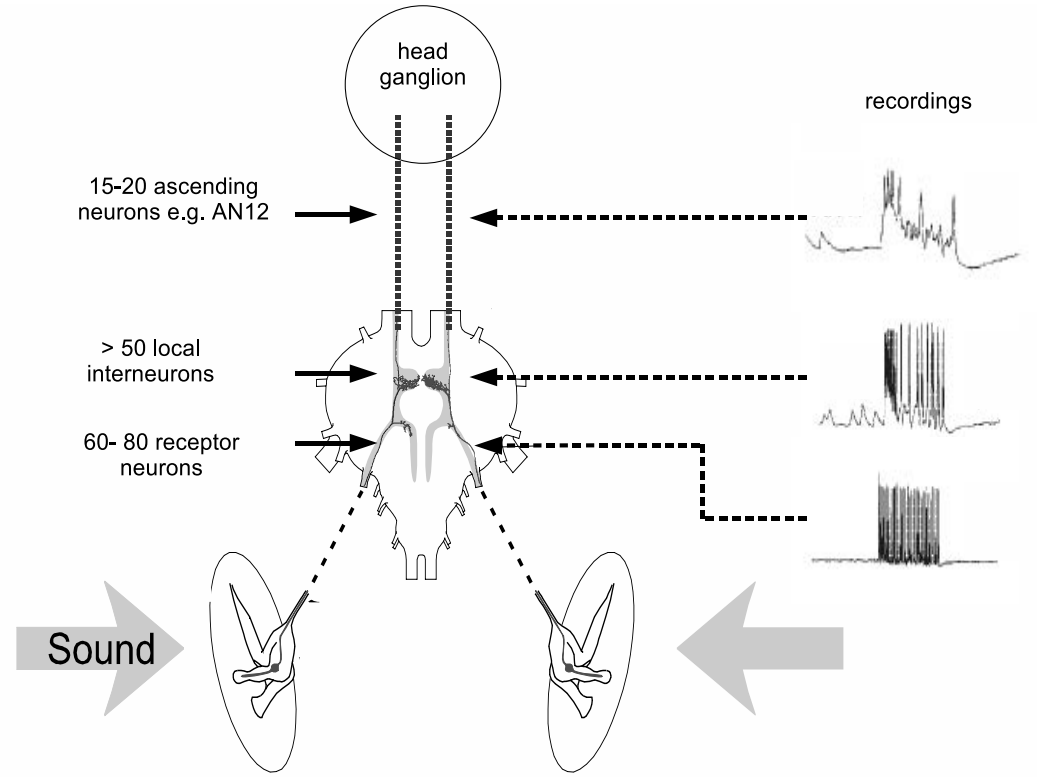


Figure 2.1: **The auditory system of grasshoppers and locusts.** Sound impinges on the two tympana where the receptor neurons translate the sound into neural activity which is forwarded to the metathoracic ganglion. Ascending neurons transmit information upwards to the head ganglion. Some exemplary recordings of different levels in the auditory network are shown. The scetch of the metathoracic ganglion is a courtesy from Hartmut Schütze, recordings are from Astrid Vogel (**?**).

The metathoracic ganglion consists of four classes of interneurons, about 100 altogether. Many have been morphologically and physiologically classified (**???**). The Ascending Neurons (ANs) form a particularly important class. They have probably no direct input from receptor neurons and are the only neurons projecting into the head ganglion. Some neurons (AN1, AN2) encode directional information, whereas others (e.g. AN3, AN4, AN6, AN11, AN12) are presumably involved in pattern recognition (**?**). Because of their small number (approximately 20), this group constitutes a bottleneck for the information transmission of the auditory system (Fig. **??**).

In a behaviourally attractive song, one of the ascending neurons, the AN12 marks the beginning of each syllable with a phasic burst (**?**). This

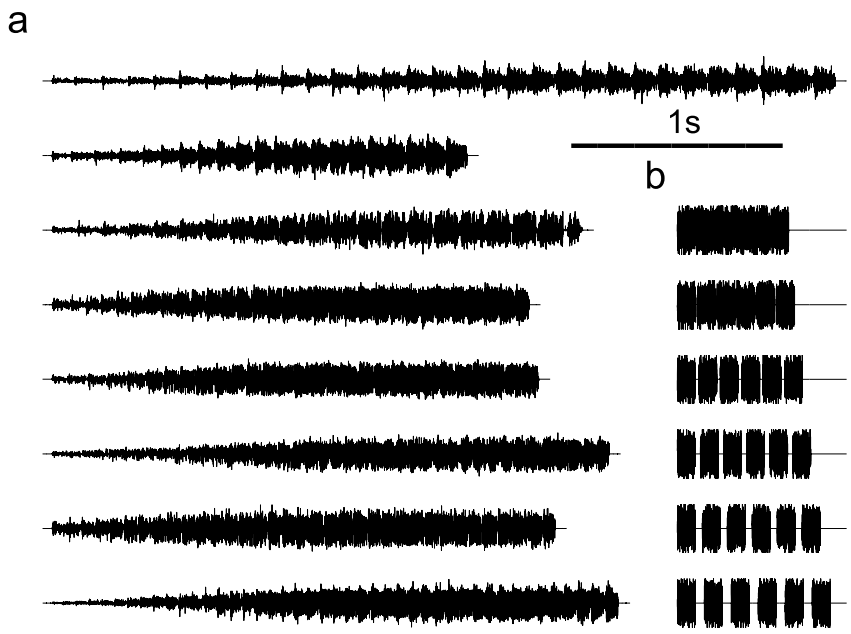


Figure 2.2: **Temporal structure of acoustic stimuli.** a) Sound pressure waves of 8 different calling songs from *Chorthippus biguttulus* males. b) Design of artificial songs consisting of blockstimuli, rectangularly modulated noise.

study also suggests that the number of spikes per syllable is positively correlated with increasing pause duration. The burst is preceded by an inhibitory post-synaptic potential (**?**). In general, AN12 is the most reliable neuron influenced by the syllable-pause structure and may account for part of the behavioural response. Another ascending neuron, the AN6, fires tonically in presence of syllables (**?**). The AN3 and AN4 respond in a phasic tonic manner to stimuli and, possibly, they encode onset steepness (**?**) and are involved in another behaviourally relevant process, gap detection (**?**).

### Grasshopper behavior

On the behavioural level, grasshoppers of the group *Acrididae* rely on species-specific song recognition (**??**) and sound localization (**?**) for successful mating. Both constitute difficult computational tasks which have to be accomplished by the auditory system. In this thesis, we focus on song recognition

and discrimination. Song recognition requires the decomposition of a stimulus into its constituents, possibly in both the temporal and frequency domain. We investigate how these decomposed constituents are efficiently encoded in later steps of auditory processing.

What are the specific features of grasshopper communication signals? Males of many grasshopper species produce songs by rubbing their hindlegs against their fore-wings. In these songs (Fig. **??**a), syllables are followed by pauses, i.e., periods of high and low amplitude modulations, respectively. Interestingly, the behavioural response depends mainly on the ratio of syllable to pause length (**?**). If this ratio is kept constant, the absolute length of one song-unit (syllable plus pause) can vary more than threefold without changing the behavioural response of the female. We analyze such *time-scale* invariant song recognition exemplifying a particular computational task that needs to be solved by grasshoppers. Gap detection forms another example: Male grasshoppers, with one hindleg missing, produce songs with gaps of at most a few milliseconds within the syllables (**?**). Females are able to detect those gaps and recognize them as an indicator of reduced fitness.

## 2.2 Encoding pause duration by intraburst spike count[1]

We analyze recordings form one specific ascending neuron in the metathoracic ganglion, the AN12 neuron (Fig. **??**), in n=6 individuals (*Chorthippus biguttulus*, n=3 and *Locusta migratoria*, n=3). Unless stated otherwise, data from *Ch. biguttulus* are shown. However, the morphological and physiological characteristics are almost identical in both grasshopper species (**??**), indicating a highly conserved functional role. Test stimuli are natural communication signals that are rhythmically structured into syllables and pauses (Fig.Â **??**a+**??**a) and artifical model songs (Fig. **??**b). Syllable and pause durations have behavioral significance as suggested by behavioral experiments with artificial stimuli (Fig.Â **??**b).

The AN12 neuron generates burst-like discharge patterns when stimulated by the amplitude-modulated sound patterns of grasshopper calling songs (Fig.Â **??**c + d). The intra-burst spike count (IBSC), i.e., number of spikes within a burst, is highly reproducible from trial to trial but varies from syllable to syllable (Fig. **??**a). Reflecting the different time-courses of different songs (Fig. **??**a), each song thus results in a particular sequence

[1]This section is mostly based on a manuscript that is going to be submitted. Detailed methods can be found in Appendix A.

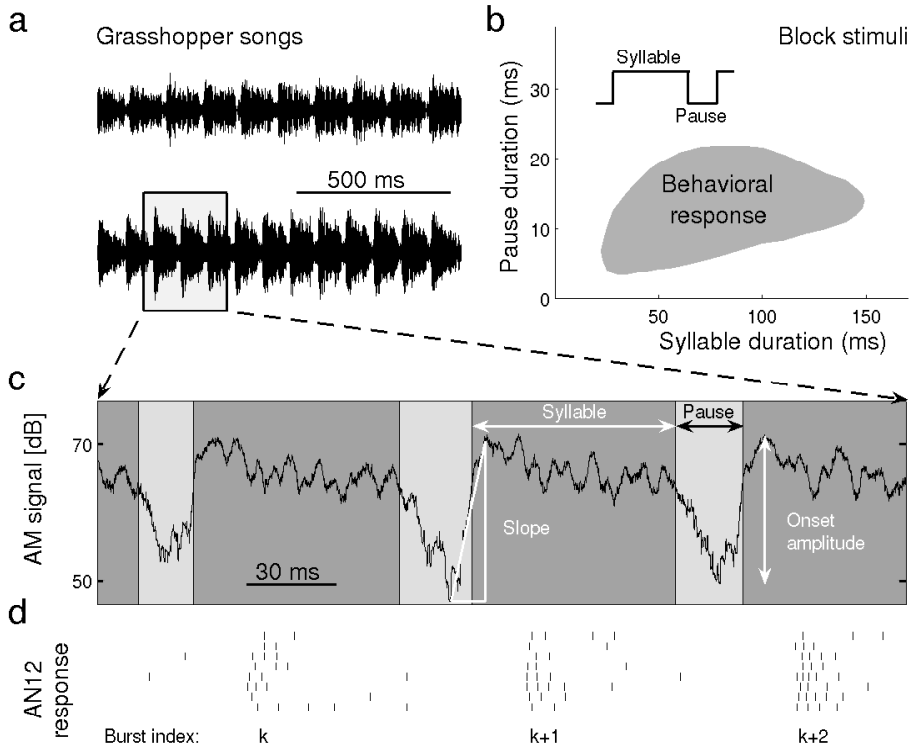


Figure 2.3: **Burst response contains information on grasshopper songs.** a) 2 examplary male grasshopper songs, consisting of a sequence of alternating syllables and pauses, episodes of loud and quiet amplitude respectively. b) Female grasshoppers respond to a range of syllable and pause durations as tested with artifical block stimuli; gray area: 1 animal at 20% positive response level. Adapted with permission from (**?**). c) Amplitude modulation (AM) signal, enlargement of song in (a). d) Spike train response of AN12, 8 repetitions. Bursts mark onset of syllables with 12 ms latency.

of IBSCs (Fig. **??**a). This signature can be used to discriminate amongst songs. For example, for a sample with eight songs from one species, each burst carries enough information to assign 40% of the responses to the correct song, using the IBSC only (Fig. **??**b). Accumulated over time, a 90%-hit rate is reached after 12 bursts, or about one second. The mutual information between spike train and song identity increases similar to the probability of correct classification, approaching the maximum of 3 bits (Fig. **??**c). This astounding discrimination performance is similar to that of grasshopper receptor neurons (Machens et al. 2003) although AN12 neurons have a far

lower overall firing rate and their exact spike timing has been neglected for the present analysis. As IBSC allows one to discriminate songs even from the same species, this measure must contain useful information about the detailed song structure.

What are the relevant features of the stimulus by which spike count within a burst is determined? To answer this question, we construct the burst-triggered average (BTA), the average stimulus preceding a burst with specific spike count (Fig **??**). A shallow peak in stimulus intensity is sufficient to elicit 1 spike. For two and more spikes a sharper stimulus peak, interpreted as a syllable onset, is preceded by a period of relative quietness. Systematically, the spike count is higher when the period of relative quietness is longer and deeper. However, neither the slope nor the relative onset amplitude do have a systematic influence on the spike count. To quantify these observations, we correlated the spike count within a burst using different measures: a) quietness period, b) relative onset amplitude, c) total period duration, d) minimal absolute amplitude, and e) slope of the syllable onset.

Most of the spike count variance is explained by the preceding 'pause' in each cell: $69 \pm 15\%$ of the variance given the external noise ($p < 10^{-5}$). For two animals, the correlation is depicted in Fig **??**a-b. The correlation is robust to changes in the amplitude level of pause duration measurement (Fig **??**c). The distributions of spike counts and pause durations have comparable shape (Fig **??**d+f). Only the onset amplitude ($27 \pm 8$, $p < 10^{-5}$) and the preceding minimal amplitude ($20 \pm 9$, $p < 10^{-5}$) can account for some variance in all cells (e.g., one cell in Fig **??**d) but have low semi partial correlations. The other factors, including the slope, are not relevant (Fig **??**a-c). Alltogether, the 5 measures can explain $77 \pm 15\%$ of the variance that is caused by stimulus statistics ($p < 10^{-5}$).

Can we interpret the doubling of a spike count as a doubling of the pause duration? This is true if the relation between spike count and pause duration can be fitted by a line through the origin. We find a systematic but small deviation from this hypothesis. The y-axis intercept lies at a pause duration of $-1.1 \pm 2.1$ ms. As the slope is at $14.6 \pm 2.5$ ms per spike count, the deviation from the 'line-through-origin' hypothesis does not exceed the level of noise in encoding accuracy, i.e., the interquartile range of pause duration at any spike count. As an illustration, the correlation is fitted by a dotted line in Fig **??**a+b. In conclusion, we can regard the observed spike-count pause-duration curve as a good approximation of a line through the origin.

To ascertain that pause duration correlates with IBSC under a variety of stimulus conditions, a second set of experiments was carried out (Ch. biguttulus, = 9), in which the pause duration in artificial song was varied systematically. Here, the 'pause' is defined as the distance between two

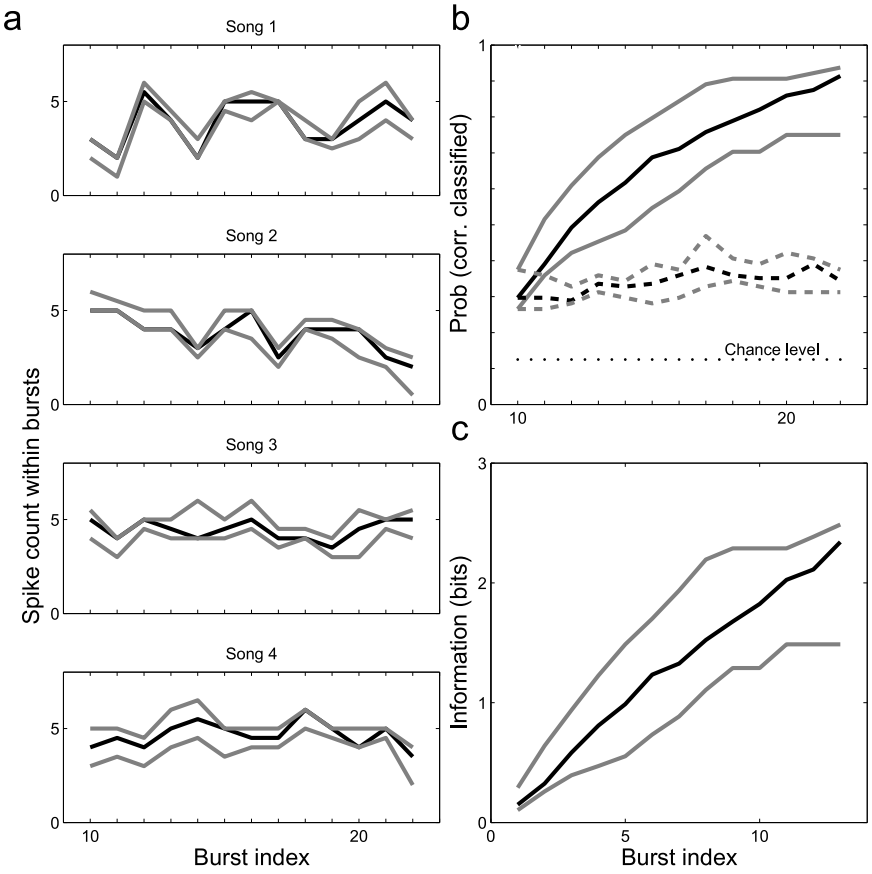


Figure 2.4: **Spike count within bursts is related to parameters in mating songs.** a) Response of an AN12 to 4 songs. The average number of spikes within a burst is plotted against the position of that burst within the total response (burst index). Gray lines depict upper and lower quartiles. b) Based on the spike count within bursts, individual spike trains are assigned to that song out of 8 songs which produces the most similar neural response. The dotted line indicates correct classification based on individual burst events alone. The solid line indicates correct classification cumulating over previous burst events. Songs can be assigned correctly with probability $> 0.9$ after 12 bursts. Hence, the spike count within bursts alone is sufficient for discrimination. c) Information about song identity as a function of the number of bursts. This is a strict lower bound as only mutual information about correct/incorrect classification was used.

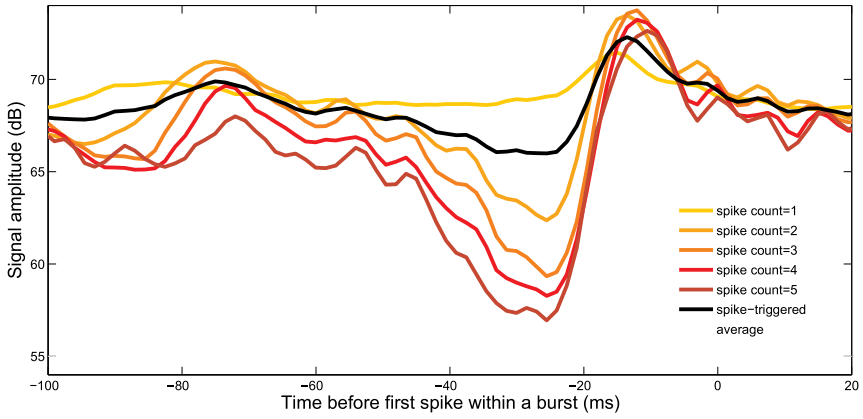


Figure 2.5: **The relationship of intraburst spike count of one AN12 cell to the stimulus.** Burst-triggered average (BTA) for natural songs: the average stimulus preceding a burst with given intraburst spike count. Higher spike count covaries with longer and deeper pauses preceding the spike count.

subsequent block stimuli (Fig. **??**b). This allows to test the relation between spike count and pause duration under controlled conditions excluding the influence of other song features. In response to such block stimuli neurons burst at the beginning of each block. We found that the IBSC is linearly correlated with the pause duration at high significance (Fig. **??**e). Over 9 cells, the correlation coefficient is $r = 0.87 \pm 0.09$ at significance levels between $p < 10^{-4}$ and $p < 10^{-17}$. The preceding syllable duration, and hence, the total period duration, does not contribute significantly to the spike count. In two cells, we tested the role of different overall amplitude levels but no clear effect could be identified.

Could the intra-burst spike count be used to reliably transmit information about the rhythmic structure of the natural calling songs or does the trial-to-trial variability blur the IBSC signal to strongly? To investigate this behaviorally relevant question, we calculated the mutual information between IBSC and preceding pause duration, using the adaptive direct method (**?**). Our data show that, on average, the IBSC transmits $0.49 \pm 0.24$ bits about the preceding pause duration. Hence a single burst would not convey sufficient information for a binary decision. However, as the trial-to-trial IBSC variability is only weakly correlated from burst to burst (correlated with $p < 0.05$ only if all cells are pooled together but not individually, turning point test **?**, for details see chapter 4), groups of subsequent bursts could be

used for transmitting information about average pause durations.

## 2.3 Why bursts?

Bursts have been described in a plethora of neural systems. Do they have properties that make them qualitatively distinct from single spikes? We would like to put our insight from the burst coding in the grasshopper neural system into a broader perspective and review burst coding in sensory systems.

Let us define bursts phenomenologically as clusters of spikes. They can be identified in interspike interval distributions that are separated into two parts: short interspike intervals for action potentials within bursts, and longer interspike intervals accounting for single spikes or interburst intervals. Bursting cells can be classified A) in electrophysiology according to their observed discharge pattern (**?**) B) in biophysics according to their burst generation mechanism (**?**) and C) in dynamical system analysis according to their dependence on parameters such as applied current (**?**). There is always an external activator, either an environmental stimulus or the state of the surrounding network. As we will see in the next chapter, a burst can be shaped by the combination of fast and slow dynamics. In this review, we chose to emphasize three aspects of bursting cells in sensory system: Feature detection and encoding, information transmission, and output performance.

### Information Encoding by Bursts

The relevance for information transmission in sensory systems has been postulated for thalamic neurons for a long time (**??**). Bursts may encode the same information as single spikes but at higher signal-to-noise ratio; or qualitatively different information; or may (additionally) be involved in extraction of behaviorally relevent features; and intraburst properties can be used for graded codes.

First, we consider examples where bursts contain more information than single spikes. In the primary auditory cortex, the carrier frequency of sound stimuli can be estimated with increasing accuracy if bursts with higher spike count are used for decoding (**?**). In a similar spirit, Livingstone et al. (1996) asked which firing pattern optimally encodes visual information in V1 in awake monkeys. The restriction to high-frequency discharges of two or more spikes (bursts) allowed a much better stimulus reconstruction of the visual input than the reconstruction from the ensemble of all spikes. This implies

that bursts can be explicitly used for efficiently encoding and decoding stimuli.

Second, bursts may contain qualitatively different information than single spikes. Place cells in the hippocampus represent the animal's location in a given environment. In such neurons, the rate of bursts depends strongly on orientation whereas the rate of single spikes has almost no dependance on orientation (**?**). In complex cells of V1, single spikes are correlated with the contrast of the stimulus whereas the clusters of spikes are tuned for spatial frequency and orientation (**?**). Further studies also indicate that bursts can selectively encode some stimulus features whereas others are represented by single spikes (**???**).

Third, bursts may also be specifically involved in the detection of behaviourally relevant events. In LGN relay cells, visually evoked bursts occur primarily at the onset of fixation (**?**). These bursts can be regarded as *wake-up calls* indicating the presence of a new kind of stimulus (**??**). In the subsequent tonic state of relay cells, coding is linear and precise stimulus reconstruction at cortical level is enabled. In weakly electric fish, the stimulus can be recovered accurately from primary afferent spike trains. The performance of downstream pyramidal neurons in encoding stimulus time courses is significantly worse than in receptor cells. However, pyramidal cells specialized in upstrokes and downstrokes, respectively, of electric field amplitudes, indicate their corresponding event by firing bursts (**?**).

Forth, intraburst properties can be used for graded codes. Modeling a pyramidal cell of the weakly electric fish, it was demonstrated that bursts occur preferentially on the increasing slope of the input current (**?**). Furthermore, within this computational model the burst duration encodes the magnitude of the slope. Along this line, it has been shown that burst interspike intervals in pyramidal cells are correlated with amplitude and slope of stimulus upstrokes (**?**). This code is reliable and precise and can be used to discriminate signals. Finally, burst duration also encodes the optimality of a stimulus in the striate cortex (**??**) and primary auditory cortex (**?**). The optimality of a stimulus tells us to what degree the stimulus fits to the tuning properties, e.g., the preferred stimulus orientation, of the neuron.

**Cellular Mechanisms underlying Burst Generation.** In many cases, intrinsic cell properties are responsible for burst generation. In *intrinsic bursters* the first spike in the burst is caused by the stimulation, but the subsequent spikes are generated autonomously due to the intrinsic properties of the cell. Sometimes the properties of specific neurons – such as the so-called chattering neurons in cat neocortex (**?**) – result in autonomous firing even without initiation of an external stimulus. Possible mechanisms of intrinsic bursting may be based on slow inward currents (**?**), $Ca^{2+}$ spikes (**?**), Backpropagation (**?**) or NMDA channels (**?**). In particular, specific $Ca^{2+}$ and cation currents can be activated (or deinactivated) at hyperpolarized potentials. Hence, bursts can occur as rebounds after release from inhibition. However, only little is known about pharmacological properties in the metathoracic ganglion of grasshoppers, but see Sokoliuk et al. (1989). Hence, modeling has to be based on another approach. Our phenemenological model (chapter 4) will demonstrate that AN12 cell behaviour can be explained by the specific temporal input distribution and no intrinsic-bursting mechanism has to be assumed.

### Bursts Increase the Reliability of Information Transmission

Single spikes are not necessarily reliably transmitted at synapses. Measuring the excitatory postsynaptic current (EPSC) elicited by stimulation of single presynaptic neurons shows that transmission probability is less than one (**?**). In fact, in hippocampal CA1 synapses the majority of synapses has transmission probability less than 0.1 (**?**). However, spike-time dependent facilitation properties of synapses allow the increase of transmission probability and, hence, synapses are thought to detect temporal firing patterns on short time scales. In fact, it has been argued that bursts are the optimal input in this regard (**??**).

More detailed studies of paired pulses at single synapses of hippocampal cells have revealed that facilitation occurs only if the first spike fails to release vesicles. If release occurs on the first spike, the transmission of the second spike is depressed (**?**). This finding has led Lisman to postulate that every burst may cause the same integrated response independent of spike count within bursts - *bursts as a unit of information* (**?**): Consecutive spikes induce exactly one transmission event, whereas single spikes are filtered out. Taking this hypothesis to the limit, bursts as a unit of information imply the

irrelevance of single spikes in information transmission. Single spikes may be regarded as noise then. This is clearly not always the case. Nonetheless, synaptic processes that cause spike-time dependent facilitation and depression are significant factors in determining the neural code (**???**), and thus bursts – as factors causing facilitation and depression – are important. A variant of this hypothesis focusses on intraburst interspike intervals (IBISI). Cells with certain IBISI may communicate selectively with those postsynaptic cells that resonate with the associated intraburst spike frequency due to synaptic facilitation and depression (**?**).

**Synaptic plasticity induced by bursts.** The transmission of bursts to downstream neurons may interact with synaptic facilitation and depression. Whereas facilition is due to a range of presynaptic processes involving $Ca^{2+}$-dependent mechanisms (**???**), depression is thought to be caused by saturation and desensitization postsynaptically (**?**). Furthermore, postsynaptic bursts paired with presynaptic activity can induce long-term potentation in excitatory synapses (**?**).

### Output performance

Neurons are commonly read out by subsequent neurons. Plausibly, not all kinds of codes can be readout by a specific decoding upstream neuron. Hence, in order to have an upstream effect information must first be transformed from one coding space into another coding space that can be *understood* by the subsequent neuron (**?**). Here, we want to point out that a burst code – either as a unit of information or as a graded code – is particularly suited for forthright readout, e.g., by temporal integration.

Bursts, thence, are a particularly suitable candidate code for the interaction between input and output. That has already been demonstrated in the cricket auditory system where interneurons, on the one hand encode salient stimulus features, on the other hand predict behavioral responses (**?**). Older studies have already highlighted that flight control in locusts is achieved via alternating patterns of bursting neurons that in turn are modified by sensory input (**??**). Similarly, in the crab, gastro-pyloric receptors can modify the bursting of cells in the stomatogastric ganglion generating rhythmic motor patterns (**?**).

Along the same line, bursts occur much more frequently when driven by natural scenes than when driven by white noise, as demonstrated in the

electroweak fish (**?**) and in the mammal visual system (**??**). Hence, it can be argued that a burst code is activated when potentially *relevant* information has to be forwarded. A mechanistic reason for this phenomenon is that bursts are activated by modulations on slow time scales when there is sufficient time for hyperpolarization between bursts (**?**). In comparison, white noise is dominated by high frequency components impeding hyperpolarization.

## 2.4 Discussion

Our review shows that bursts can play a decisive role in encoding information, transmitting information and giving rise to an effective output. In some studies concerned with information transmission (**??**), the fine structure of bursts conveys no specific stimulus-related information. In contrast, in the electroweak fish it has been shown that intraburst interspike intervals carry information on amplitude and slope of stimulus upstrokes (**?**). Modelling studies indicate that the spike count within bursts or a similar measure, the burst duration, can encode the amplitude slope of sensory signals (**?**). Our investigation of the bursting interneuron in the grasshopper auditory system demonstrates that also intraburst spike count can encode information about stimulus features. In particular, we have presented evidence that the AN12 encodes the pause duration between subsequent syllables by spike count within bursts (Fig **??**a-b,e), but not the slope (Fig **??**a).

Burst duration, spike count within bursts and burst ISIs can all be used to encode specific information. Their presence signals that something important is happening. Information transmission via bursts is reliable. They have sufficient power to change the subsequent output, e.g., the cortical (perceptual) state in mammals or the behavioral response in invertebrates. Intrinsic bursting mechanisms allow flexibility in burst coding: it may be sensitive to context information. Bursts, including distributed bursts, may be regarded as the focal point between coding and effective change in behavioral states.

Crucially, these graded burst codes (electroweak fish, grasshopper) display an interplay of different codes. The existence of bursts signals the presence of a pause in a binary fashion, as a single spike would do. In fact, the interburst interval represents the period duration of the communication signal. However, additionally the intraburst spike count codes for the pause duration, thus, constituting an additional graded code. By multiplexing both codes into one neural event a joint readout is enabled. Furthermore, the AN12 investigation also shows that a temporal feature with 40 ms duration can be compressed into a code of 4 ms duration.

Bursts are also suited to cause significant change in the state of upstream

neurons or – for that matter – behavioral output. In thalamic relay cells, bursts occur more often when driven with natural occuring statistics (**?**), indicating that bursts are suited to transmit *relevant* information. Bursting motor neurons are substantially influenced by sensory input (**?**). Already in the sensory system, bursts can predict behavioral responses (**?**). The burst code in the AN12 neuron also encodes a behaviorally relevant temporal signal, the pause duration. In the next chapter, we will investigate how the spike train of this bursting interneuron is decoded such that an even higher-level property of the communication signal is extracted.

### Summary and Outlook:

The grasshopper auditory system is the model system of our study. As a particular asset of this system, electrophysiological, anatomical and behavioral properties have been characterized in detail. We focus on the processing of temporal patterns of grasshopper communication signals in a bursting interneuron. We show that intraburst spike count encodes a specific feature of the communication signal – pause duration. The role of bursts in sensory system is discussed at full length. In the next chapter, we suggest a plausible read-out of this bursting interneuron.

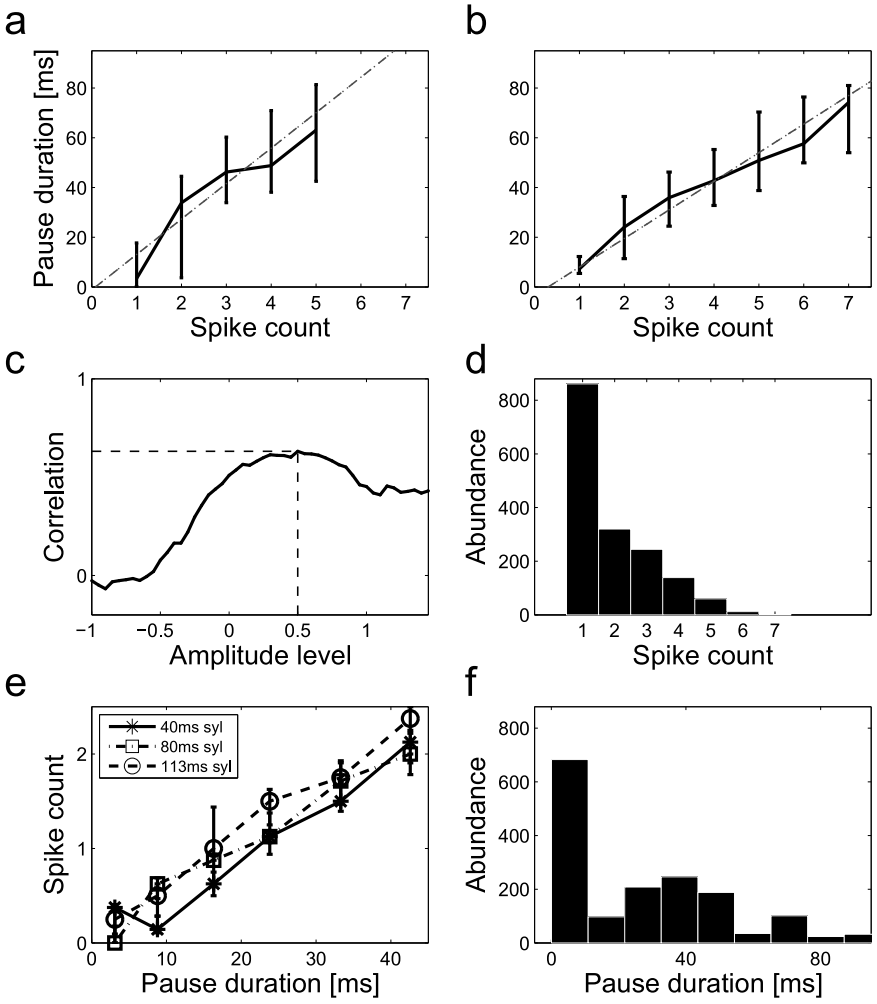


Figure 2.6: **Spike count scales with pause duration.** a+b) Distribution of the pause duration preceding a burst as a function of the burst spike count in two cells. The correlation can be approximated by a line through the origin. c) The amplitude level for measuring pause duration is determined by optimizing for the correlation between pause duration and amplitude levels. The resulting correlation is not critically dependent on the amplitude level setting: Varying the amplitude level which defines pause duration shows that correlation between pause duration and spike count within burst is robust with respect to level setting. d) Overall distribution of spike counts. e) In artificial model songs consisting of block stimuli (see stimulus of Fig **??**b), the pause duration is systematically varied and spike count is measured over repeated trials, demonstrating the correlation under controlled conditions. f) Overall distribution of pause durations in natural songs.

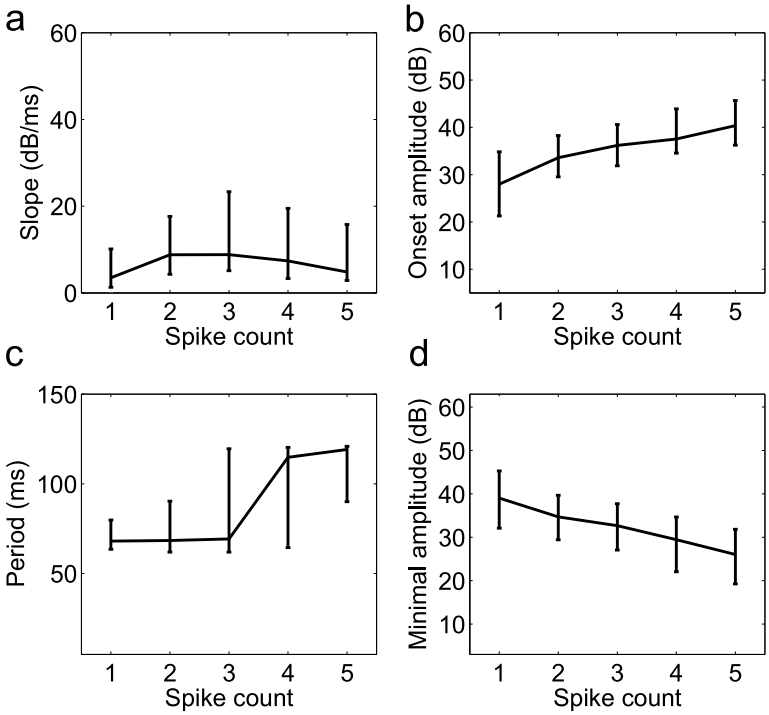


Figure 2.7: **The relationship of spike count within bursts of one AN12 cell to other stimulus parameters.** a) The distribution of the onset slope as a function of the corresponding burst spike count in AN12. Indicated are median, lower and upper quartile values. b) The distribution of the relative onset amplitude preceding a burst as a function of the burst spike count. c) The distribution of period durations preceding a burst as a function of the burst spike count. d) Distribution of the absolute minimum amplitude within a pause as a function of the subsequent burst spike count. A deeper pause leads to higher spike count.

# Chapter 3

# Time-scale invariant decoding

Object recognition relies on the extraction of stimulus attributes that are invariant under natural variations of the sensory input. Such stimulus variations include the size, orientation or contrast of a visual object (**???**), the strength of an odor (**??**), and the amplitude, pitch or duration of a sound signal (**????**). A particular challenge arises when time-scale invariant features of an acoustic signal are to be extracted as this computation involves ratios of temporal quantities. In other words, stretched or compressed temporal sequences - such as GOAL! or GGGOOOAAALLL!!! - need to be classified as equal. To calculate the relative duration of two specific sound pattern within a longer stimulus the respective duration of both components need to be measured and their ratio be computed. Using grasshopper communication (**?**) as a model, we here demonstrate that this seemingly difficult task can be solved in real time by a small neural system. As shown in the preceding chapter, an auditory interneuron generates bursts of action potentials in response to natural calling songs and simplified artificial stimuli that mimic the rhythmic syllable-pause structure of grasshopper calls. The recorded in-vivo data show that bursts are preferentially triggered at syllable onset and that the intra-burst spike count scales linearly with the duration of the preceding pause. Integrated over a fixed time window, the total spike count thus contains information about the syllable-to-pause ratio of the presented song. Since this ratio is species specific, the information has a high behavioral value. The underlying neural coding strategy is robust in that it does not require any division of time-dependent quantities. The encoded time-scale invariant information can be read out easily by down-stream neurons.

## 3.1 The puzzle of temporal sequence identification

Many animals use acoustic communication signals to find conspecific mates and judge their reproductive fitness (**?**). How does a sensory system recognize these behaviorally important spatial and temporal signal patterns? Rather than mapping external stimuli one-to-one into a neuronal response, sensory systems selectively extract relevant information (**??**). In the auditory system, for example, it is well understood how interaural time differences can be used to identify sound location with high resolution (**??**). However, neural systems that represent temporal patterns are confronted with an additional challenge: information that is spread over time needs to be pooled into an explicit neuronal event, at the latest on the level of motor output. In contrast to invariance computations in the visual system (**??**), such a computation of temporal sequences is poorly understood. This encoding and decoding problem is aggravated when temporal patterns must be identified invariantly to fluctuations of the time scale. Such time-warp invariant sequence recognition is an important task for some auditory systems. In human speech, words are identified with ease even when the speaking rate is varied (**??**). For speech recognition, Hidden Markov Models can identify speech successfully but are limited in treating variable time duration in speech (**?**). In grasshoppers, poikilothermic animals, the overall temporal scale of communication signals varies strongly with changing temperature conditions. In particular, sitting in the shadow of a tree the grasshopper sings slower than when bathing in the sun. In fact, the grasshopper *Chorthippus biguttulus* produces syllable-pause sequences that vary up to 300% without impairing the stimulus recognition even when the body temperature of the receiving animal is unchanged (**?**). Previous theoretical solutions to the general time-warp problem are based on subthreshold oscillations and coincidence detection (**??**), transient synchronization (**?**), a synfire chain with two different sorts of inhibition (**?**) and maximization of the variance of input currents (**?**). Some of these models use a set of neurons, e.g., with different decay times, facilitating speech recognition at the expense of high computational costs. The grasshopper, in contrast, needs time-warp invariant stimulus recognition only to reach a binary decision when being courted with a mating song: To respond or not to respond. Considering this low information rate of the output, no computationally expensive network (in terms of number of neurons) is required. What kind of simple mechanism solves this problem?

We have already seen in Chapter 2 that for grasshoppers, communication signals play a decisive role in finding a potential mate (**??**) and are thus di-

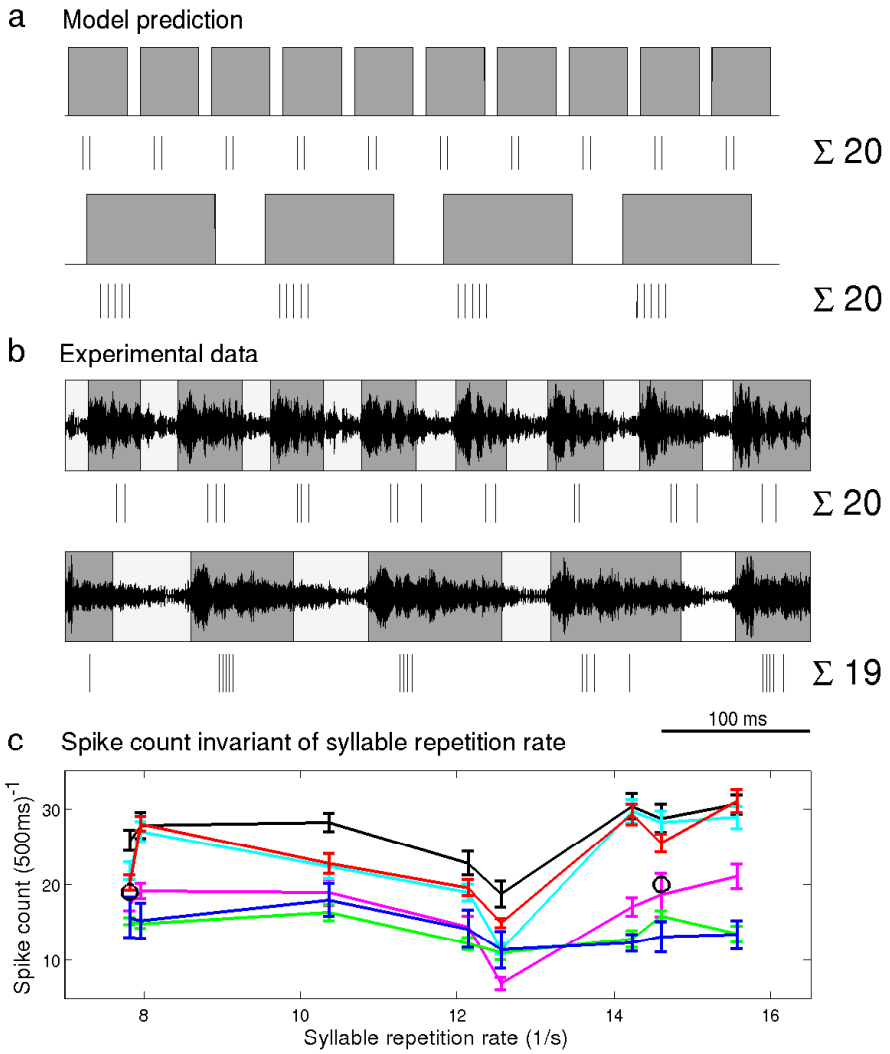


Figure 3.1: **Total spike count is time-scale invariant.** a) Scetch of 2 block stimuli. The lower sequence is obtained by scaling the upper sequence by a factor of 5/2. This results in 4 instead of 10 bursts but each burst has 5 instead of 2 spikes. These two effects compensate each other: the total spike count is constant. b) Response to two natural songs with different pause durations. c) Total spike count plotted against the number of syllables per second in 8 different songs. 6 animals are color-coded. In average, no correlation can be found (R2=0.00 $\pm$ 0.14). The indentation corresponds to two songs that cause only weak firing within the measured period. However, those two songs cause higher firing rates at the beginning of songs. The two examples of b) are marked as circles.

rectly related to reproductive success. Recall that mating songs of grasshopper males consist of alternating syllables and pauses which can have strongly varying durations and that behavioral experiments with model songs, consisting of block stimuli (interpreted as syllables) and intermediate pauses show that the ratio between syllables and pauses is more relevant for behavioral response than the absolute durations of syllables or pauses (**?**). In fact, grasshopper songs as a function of temperature are not modulated on subsegments but are globally scaled. Therefore we here investigate an instance of time-scale invariance, a subset of time-warp invariances.

## 3.2 Time-scale invariant integration of the bursting interneuron

Can the properties of the bursting interneuron AN12 and the putative pause duration encoding mechanism be understood in the context of time-scale invariant behavioral response? Time-scale invariance in this context means that the ratio between pause duration ($T_{pause}$) and period duration ($T_{per} = T_{pause} + T_{syl}$) is kept constant. Hence, a first hypothesis is that the grasshopper auditory system performs a division: $\frac{T_{pause}}{T_{per}} = \text{const}$. All necessary information for this operation is transmitted by the AN12: the spike count within bursts encodes the pause duration; the first spike within a burst encodes the time of the syllable onset and, hence, the interburst interval encodes the period. Of course, the time of the syllable onset itself has no meaning for the organism; valuable information is only in relative timing, i.e., in the interburst interval. A sophisticated read-out mechanism could calculate the ratio between pause duration and period. However, such a divisor might be very sensitive with respect to time scales and is not necessarily robust against noise in spike count.

Instead consider as a second hypothesis that female grasshoppers measure the ratio of the average pause to the average period within a time frame $T_{dec}$. This ratio depends only on the total pause duration within the time frame and can be interpreted as a moving average. Formally,

$$\frac{\widehat{pause}}{\widehat{period}} \sim \frac{1}{T_{dec}} \int_0^{T_{dec}} pause(t)dt$$

where $pause(t) = 1$ if there is a pause at $t$ and 0 otherwise and $\widehat{pause}$ and $\widehat{period}$ indicate the mean of the corresponding quantities. Such a moving average alters the division into an integration operation. The experimental observations from the previous chapter imply that the spike count within

bursts in the AN12 neuron is proportional to the preceding pause,

$$\text{intraburst spike count} \sim \text{pause duration} .$$

Hence, a plausible read-out neuron simply has to count spikes, by this, measuring the total pause time within a relevant time frame. If syllable and pause durations were multiplied by the same factor, this measure would stay constant (Fig. **??**a). Indeed, measuring spike count over a long time-window, e.g., 500 ms in response to different songs lead to similar results (Fig. **??**b). Each AN12-neuron keeps the total spike count approximately constant in response to different songs (Fig. **??**c). Hence, the measured response of AN12-neurons supports our second hypothesis that long-time integration is sufficient for time-scale invariant decoding.

How large is the deviation in the spike count? We measured the maximal deviation for a given song from the total mean average spike count over all songs, measured as the percentage of the mean spike count, obtaining for the 6 cells (Fig. **??**c) $38 \pm 14\%$ deviation; allowing for one outlier in each cell (one song eliciting negative response), one obtains $22 \pm 5\%$ deviation. Applying this for natural songs, we assume that the integration time is related to the minimal song duration eliciting behavioral responses (**?**), i.e., about one second.

Temporal integration over such longer time scales is well known, for example, in the electric fish (**?**). The plausibility of the proposed counting mechanism for read out has been demonstrated in frogs that, similarly to grasshoppers, also communicate with patterns of pulses and quiet intervals. Here, specific neurons in the frog's auditory midbrain integrate the number of acoustic pulses and respond only if a minimal number of appropriate pulses and interpulse intervals arrive (**??**). Furthermore, neurons in auditory cortex of mammals integrate over a variety of timescales (10 ms, 100 ms, 1 s) (**?**). Models of invariance computation suggest that high-level invariances are detected by successive spatial integration over low-level features (**??**). Our study relates both results by indicating that invariant recognition of temporal patterns can be achieved by successive integration over longer time-scales, i.e., pause integration over 10-100 ms and song integration over 1 s.

## 3.3 A mechanism for human speech recognition?

The integration of a temporal feature detector leads only to time-scale invariance for periodic signals. In principle, such an integration could be used

to identify piano warblers or other periodic music. It does not provide a straight-forward mechanism for time-warp invariant representation of human speech. In contrast to grasshopper songs, human speech recognition is a complex process based on the evaluation of spectral components and a high number of temporal features. However, there is increasing evidence that the peripheral human auditory system effectively utilizes syllable-sized time-spans ($\sim$ 200 ms) of the audio signal (**?**), indicating the possibility that integration participates in time-warp invariant speech recognition as well.

### Summary and Outlook:

Time-scale invariant recognition of communication signals is an important behavioral task for grasshoppers. Here, we suggest a forthright read-out of the bursting interneuron AN12 that is invariant to overall scaling in time. Instead of performing a division, the hypothetical postsynaptic neuron could integrate the spike count. Total spike count over a behavioral time window is an invariant of time-scaling. In chapter 5, we will incorporate this model into a broader framework that can elucidate song recognition.

# Chapter 4

# Fast excitation and slow inhibition: a canonical circuit for auditory processing?

In chapter 2, we have seen that spike count within bursts in the AN12 neuron encodes preceding pause duration. But what is the underlying circuit causing this bursting response? In the first part of this chapter, we explore a putative model based on differential equations that reproduces the neural bursting response to grasshopper communication signals. The differential equations can be interpreted in terms of a parallel neural circuit consisting of a fast excitatory and a slow inhibitory input channel. In the second part of this chapter, we compare this neural circuit, termed FexSin model, with the auditory system of songbirds. Though the latter system is much more complex and can discriminate between a large variety of bird songs, a similar kind of circuit can simulate neural response in the auditory forebrain phenomenologically and under a variety of noise conditions. Furthermore, fast excititation and slow inhibition may be responsible for precise firing in the auditory cortex of mammals. The seeming abundance of this circuit is not unexpected as 1) the operation is very basic in nature and 2) can be regarded as stimulus-induced adaptation and, hence, as a strategy to efficiently encode information under noisy conditions.

## 4.1 The FexSin circuit

We introduce the fast excitation - slow inhibition (FexSin) circuit by modeling the response of the AN12 neuron to natural communication signals. The goal is to optimize spike train similarity between observed neural data and model

data.

First, the signal input is processed by an adaptation current in the receptor neurons. The resulting current is interpreted as an excitatory input for the AN12 neuron. The inhibitory channel consists of an additional first-order low-pass filter of this receptor response. The AN12 neuron is modeled as an integrate-and-fire neuron with both excitatory and inhibitory input. In the first step, adaptation $a(t)$ current is a first oder low-pass filter of the stimulus $s(t)$ with time constant $\tau_{rec}$ (Equ. **??**). This adaptation current is then substracted from the sound stimulus $s(t)$ with weigthing factor $A_{rec}$, constituting the effective receptor response $r_{rec}(t)$, (Equ. **??**). $A_{rec} \in [0, 1]$ can be interpreted as the relative adaptation level.

$$\tau_{rec}\frac{da(t)}{dt} = -a(t) + s(t) \tag{4.1}$$

$$r_{rec}(t) = s(t) - A_{rec}a(t) \tag{4.2}$$

The receptor response is forwarded to a putative interneuron that excites the AN12 model neuron, (Equ. **??**). The inhibitory current $r_{inh}(t)$ is modeled as a first-order low-pass filter of the receptor current (Equ. **??**).

$$r_{exc}(t) = r_{rec}(t) \tag{4.3}$$

$$\tau_{inh}\frac{dr_{inh}(t)}{dt} = -r_{inh}(t) + r_{exc}(t) \tag{4.4}$$

The effective input to the AN12 model neuron, $r_{exc}(t) - A_{inh}r_{inh}(t)$, consists of excitatory input and inhibitory input, weighted with $A_{inh}$. This effective input serves as the driving force of a leaky integrate-and-fire AN12 model neuron with membrane potential $V(t)$:

$$\tau_{RC}\frac{dV(t)}{dt} = -V(t) + r_{exc}(t) - A_{inh}r_{inh}(t)$$

The model neuron spikes whenever the voltage $V(t)$ passes a certain threshold value $V_{th}$ and $V(t)$ is reset to $V_{reset}$ for a refractory period determined by measuring the minimum interspike interval of the observed data.

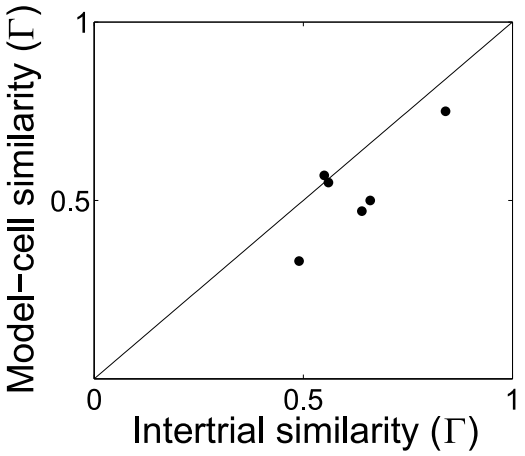


Figure 4.1: **Comparing the model quality to intertrial variability.** Γ-values measure the degree of similarity between two spike trains given the same stimulus. Γ-values of the intertrial variability are plotted against the Γ-values of the optimized model neuron. The model quality increases for neurons with lower intertrial variability, i.e., higher reliability.

The quality of the model is measured by counting the normalized input spikes. Denote with $n_{AN12}$ and $n_{model}$ the total number of spikes in the average recorded AN12-neuron spike train and model spike train, respectively, and $n_{coinc}$ the number of coincidence spikes. Then the coincidence measure is defined as: $\Gamma = \frac{2n_{coinc}}{n_{AN12}+n_{model}}$. Model parameters are chosen such that $\Gamma$ is maximized.

The average spike train is constructed across the repetitions of the recorded AN12 data: whenever there is a burst at a certain moment, the average spike count across all 8 repetitions in a time window of 2 ms is calculated; the burst time was determined by the time of the first spike; all spikes within a burst were treated as occurring at the time of the first spike. By this, spike trains are compared on basis of intraburst spike count (IBSC) only but not on the temporal fine structure within bursts. This measure is appropriate to find a model that reproduces IBSC response behavior. Bursts in the recorded AN12-data and in the model spike train are counted as being coincident when the burst times are not more than 2 ms apart, by this contributing to $n_{coinc}$. Bursts where the first spike is displaced by more than 2 ms are treated as different events. If both spike trains are identical, $\Gamma = 1$. If no coincident spikes occur, $\Gamma = 0$. As a reference value, we obtained the intertrial similarity by measuring the Γ-value between repetitions of the same song and its average spike train for each cell (range: $0.49 - 0.84$; median=0.60) in the recorded AN12-data. To calculate the quality of the model, we fitted all 5 parameters to the average spike train for 7 songs and calculated the Γ-value for the 8th song. We repeated this procedure for all songs and computed the

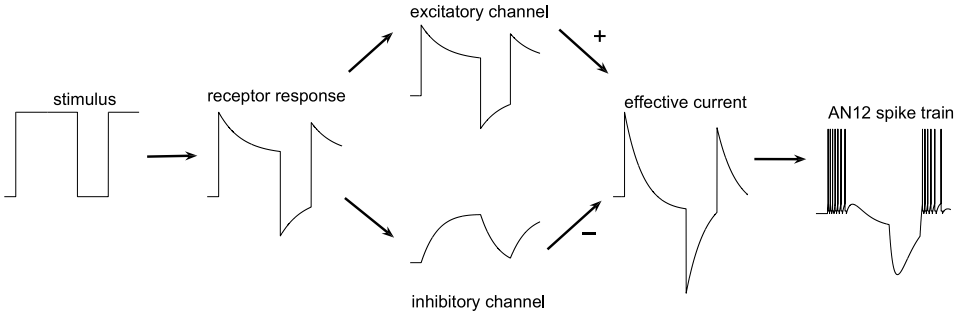


Figure 4.2: **Model simulating AN12 neuron response.** To demonstrate the circuit processing, a block stimulus is used as input. Receptor neurons are adapting ($\tau_{rec} = 30$ ms), emphasizing the first part of the syllable. The excitatory channel forwards this signal to the AN12, whereas the inhibitory channel low-pass filters the signal. Both channels are summed up to form the effective current which drives the leaky integrate & fire neuron.

mean Γ-value. Each cell was fitted with a parameter set. For the 6 cells, we obtained: range: $\Gamma = 0.33 - 0.75$; median: $\Gamma = 0.53$. That is at most 30% away from the reference Γ-value in each case. Γ-values are plotted against each other in Fig. (**??**).

## 4.2 Dynamics driving the bursting interneuron

By encoding the preceding pause duration, AN12 responses depend on the stimulus history. We suggest two hypotheses that could explain the dynamics of this bursting interneuron. The first hypothesis is that cell-intrinsic mechanisms like adaptation currents (**?**) play a role. The second hypothesis is that a preceding circuit causes the phasic response.

Let us test the first hypothesis. If there were spike-firing dependent adaptation on time scales longer than syllable duration, then we would expect a negative correlation between the deviation from the mean of subsequent spike counts within bursts (**??**). Put more simply: Many spikes in one burst would imply less spikes in the next burst. Using a turning point test to investigate burst interval correlations (see text box below), we found, if at all, only a weak role of spike-driven adaptation on long time scales ($t = 80 - 150$ ms). In each cell, there is a small but not significant (at the 5% level) upward bias in turning points, indicating negative correlations. However, if we pool all cells

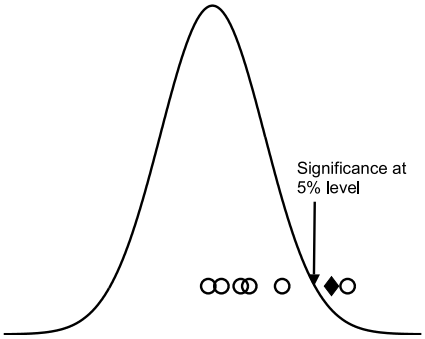


Figure 4.3: **Weak correlation between subsequent spike counts.** Plotted is the expected normalized distribution of turning points. Individual cells (circles, $n = 6$) display a slight bias having an above average number of turning points, but not significantly. However, if one pools all 6 cells together, the upward shift proves to be significant (diamond). This indicates that spike-firing dependent adaptation plays a weak role.

together, the bias is significant at the 5% level (5807 turning points, expected: $5718 \pm 77$). Crucially, spike-firing depenpent adaptation would imply scaling of IBSC with preceding period but not pause duration. In agreement with this consideration, our efforts to model AN12 response based on adaptation mechanism resulted in relatively low quality models. We conclude that firing rate dependent adaptation of the AN12 at long time scales is not decisive for AN12's bursting behavior.

**Testing for adaptive currents in AN12**. Adaptive currents lead to negative correlations of subsequent intra-burst spike counts. We used the turning point test (**?**) to find deviations from randomness in the time series of subsequent inter-burst spike counts, by this checking for negative correlations. If a time series is purely random, two third of all points are expected to be turning points. Taking boundaries into account, we expect $(2n-4)/3$ turning points in a time series of n consecutive points. For large n, turning points should be distributed as about $N(2n/3, 8n/45)$ in a random time series.

According to the second hypothesis, the architecture of the neural system is responsible for the bursting response. The data analysis showed that the driving force of AN12 excitation is the duration of the preceding quietness at syllable onset. We postulate that this phenomenon can be understood by the FexSin model introduced in the last section.

In the following, we provide the details of the model. Here, the parameters of the receptor neuron are fixed to $A_{rec} = 0.5$ and $\tau_{rec} = 30$ ms which

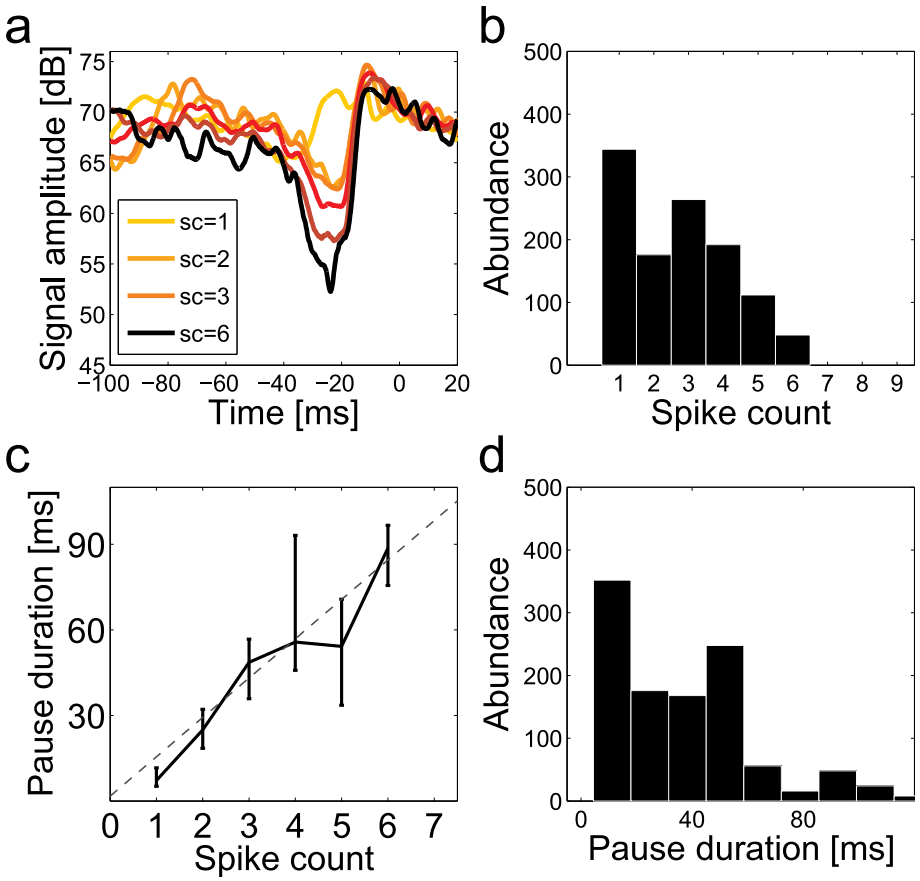


Figure 4.4: **Response behavior of the model neuron.** a) Burst triggered average. b) Overall distribution of spike counts. c) Distribution of the preceding pause duration as a function of the burst spike count. d) Overall distribution of pause durations. Compare with Fig. **??** and Fig. **??**.

is within the range of observed values (**?**). Our model does not crucially depend on the choice of these parameters. The other five model parameters (the two inhibitory channel parameters $\tau_{inh}$, $A_{inh}$ and the three integrate & fire parameters $\tau_V$, $V_{th}$ and $V_{reset}$) are optimized with respect to the model quality criterion. In each cell, the obtained parameters of the model inhibitory neuron ($\tau_{inh}$, $A_{inh}$) were in the same range. Hence, we fixed the two values globally to $\tau_{inh} = 40$ ms and $A_{inh} = 1.3$, but optimized the integrate & fire parameters independently for each cell.

The model has 2 versions. The first is without receptor adaptation, the second includes receptor adaptation. Without receptor adaptation, 48% of

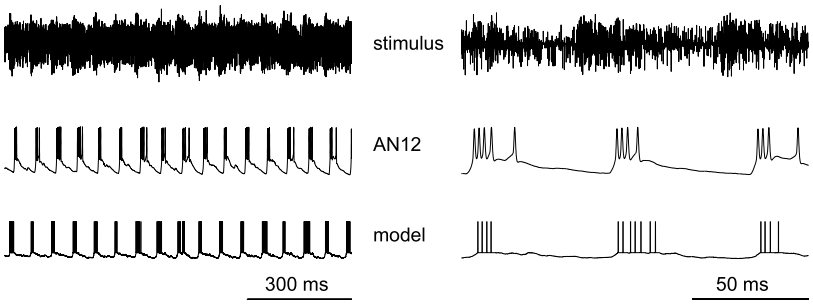


Figure 4.5: **Comparison of AN12 and model response** on two different time scales, left: 1300 ms, right: 150 ms

the spike count variance can be explained altogether (compare with 59% in the recorded data). 83% of this explained variance can be attributed to the preceding pause (69% in the recorded data). The model without adaptation fits well to recorded spike trains. However, when tested with artificial block stimuli, the model generated up to 26 spikes per burst, far more than its biological counterpart ($\sim$ 10 spikes). Unlike natural songs, block stimuli stay at the same amplitude level. Thus, the high amplitude level in the second part of the syllable leads to high spike count which would not be possible under natural song condition.

The extended model, including receptor adaptation, is summarized in Fig. **??**. This model explains AN12 spike trains under natural stimulus conditions: the spike count levels at 12 spikes for long pause durations in response to artificial block stimuli. In the extended model, 45% of the spike count variance can be explained by stimulus features, 36% alone by the preceding pause duration. In Fig. **??** we depicted the main attributes of the complete model. Fig. **??**a shows a similar burst triggered average as in the neural data. The spike count distribution (Fig. **??**b) is similar to the spike count distribution of the AN12 neuron (Fig. **??**d). The correlation between spike count and pause duration is approximately reproduced (Fig. **??**c). Note that the spike-count pause-duration correlation is optimized on different threshold levels in real and model cell, leading to a modified pause duration distribution (Fig. **??**f + **??**d). Fig. (**??**) shows the model neuron's spike train in comparison to real data, both responding in a similar manner to the natural song.

The model cannot explain all the variance of neuronal data. Furthermore, the model may not be sufficient to reproduce AN12 response under different stimulus paradigms. However, the model features a simple, i.e., low

dimensional, rationale of AN12 responses to natural and artificial songs.

To summarize: our model gives an account of the AN12 spike train with respect to two different stimulus conditions: natural mating songs and artificial block stimuli, based on the summation of fast excitation and slow inhibition. Other models may also be plausible. For example, a cation channel deinactivated at hyperpolarization could participate in producing a burst of spikes. However, as the spike count is proportional to pause duration but not period, we would still rely on a inhibitory neuron marking pause duration. Such a model, based on channel dynamics, would in fact increase the complexity of our model.

## 4.3 Neural response in the auditory forebrain of songbirds

Do other auditory systems have similar features as the suggested simple circuit for the grasshopper system - based on fast excitation and slow inhibition? As this operation is very simple and nonetheless useful for empasizing onsets we expect the answer to this question to be positive. In this section, we focus on the auditory forebrain of songbirds. The songbird auditory system is certainly much more complex than the grasshopper one. In fact, songbirds can discriminate a wide variety of complex natural sounds. Both vocal communication behavior and identified neural circuits that mediate perception are fairly well understood (**??**). It has been suggested by anatomical and physiological studies that field L is involved in signal discrimination (**??**). Field L lies between the thalamic relay nucleus and higher cortical areas such as HVc. Indeed, a spike-timing based code of some single neurons in field L can be used for stimulus discrimination matching behavioral accuracy (**?**).

For acoustic communication, sound sources of interest have to be separated from other distracting acoustic stimuli and noise sources. This problem is particularly relevant in the context of speech intelligibility. A variety of studies indicates that humans can cope with "maskers" (for a review see **?**). This poses the question how neurons in higher areas of auditory systems can reliably identify sound sources in the presence of distractors. Kamal Sen and colleagues looked at the spike train response of cells in field L of songbirds under a variety of noise conditions. In particular, three maskers with the same long-term spectral content but different short-term statistics were used. (1) Modulated noise: Broadband noise with spectral profile matching that of the average of random addition of three song motifs (so-called chorus), modulated by the envelope of a random chorus. (2) Noise: Broadband

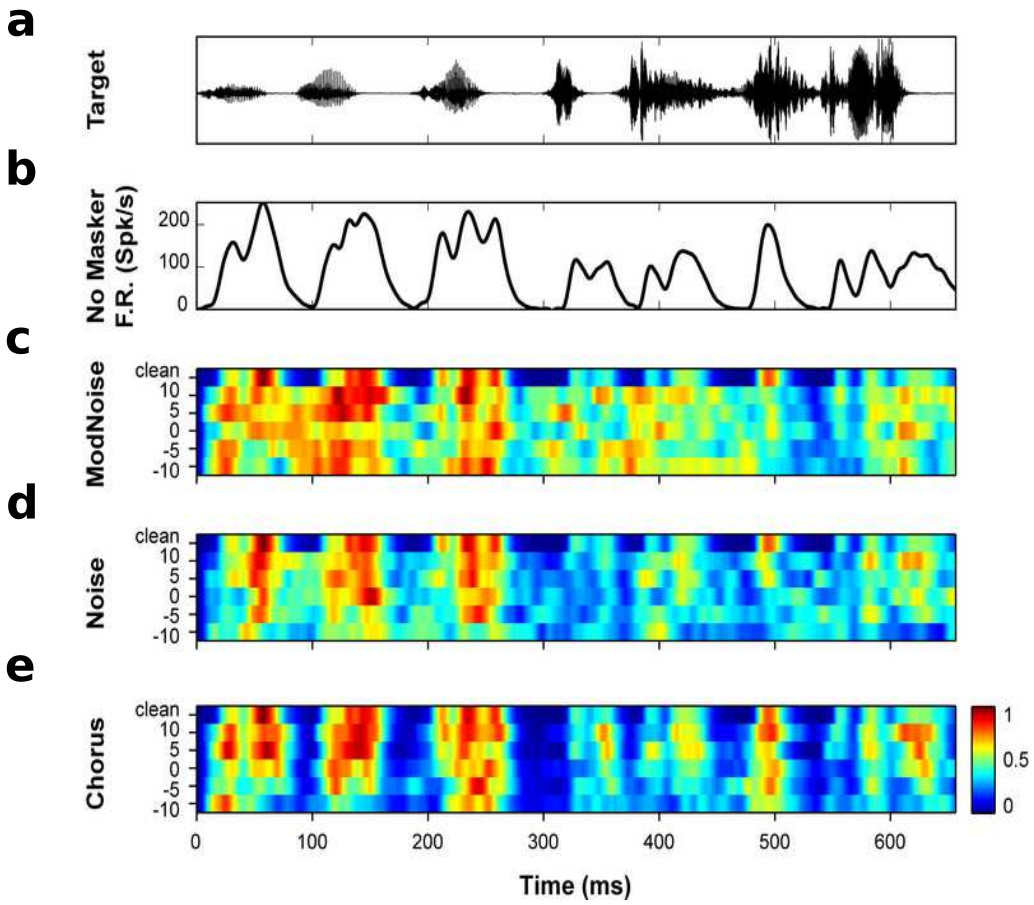


Figure 4.6: **Typical song motif of songbirds and neural response under a variety of masker conditions.** a) The pressure wave form of a target song motif. b) The neural response of a single neuron in field L of songbirds. c) The average normalized spike rate of this neuron color-coded in response to the song (first row) and to different amplitude levels of added modulated noise (row 2-6). The y-axis indicates the difference in dB between song and masker amplitude. d) Similar to c) but with unmodulated noise. e) Similar to d) but with chorus masker. This figure is adapted from a presentation by Professor Barbara Shinn-Cunningham.

noise as specified above, not modulated. (3) Chorus: Addition of three song motifs. For details of the masker generation see **?**.

A typical song motif of a songbird consists of a specific amplitude modulation (Fig. **??**a), similar to the grasshopper song. However, in difference to the latter, songbirds have a rich repertoire of different syllables. Furthermore, frequency bands vary across the song motif. A particular neuron

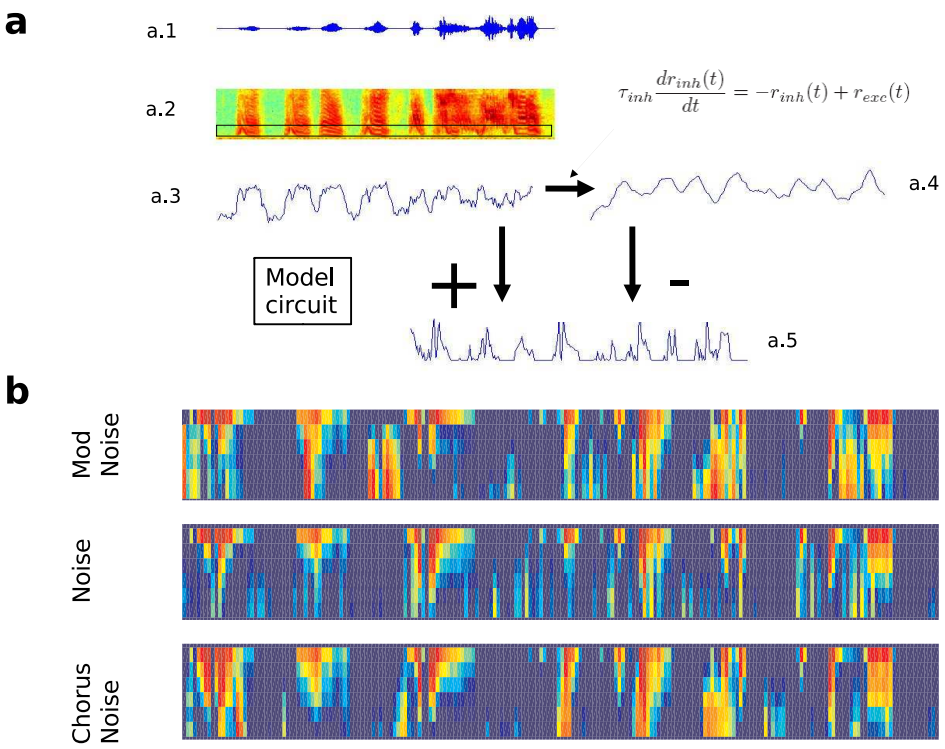


Figure 4.7: **Simple model for the field L neurons and model response.** a) The song motif (a.1) has a particular power spectrum (a.2). For the model of the narrowband neuron, the lowest 2 kHz are extracted from the power spectrum, indicated by the black frame. The average amplitude within this spectrum is translated into the instanteneous firing rate of an excitatory neuron (a.3). This firing rate is convoluted with an exponential function with $\tau_{inh}$ into the firing rate of an inhibitory neuron (a.4). The joint input of both neurons gives rise an output firing rate of the model neuron (a.5). b) Response of model neuron to different noise condition as in Fig. **??**.

in field L responds with a firing rate following the time course of the song up to a certain degree (Fig. **??**b). In fact, this neuron is a narrow-band neuron, responding only to a particular frequency band, hence, not following the amplitude modulation accurately at all times. When presented with different maskers additional to the target song motif, the neuron maintains the response profile for increasing amplitude levels of the masker. For the modulated noise condition, one observes an addition of spikes. In contrast, for the broadband noise and chorus masker condition, under conditions of high masker amplitude spike suppressions occurs for points in time where

the firing rate was elevated for the reference (clean song) condition. How can this observation be explained?

In the following, we will present a model reproducing the observations above[1]. The model will be similar to that one of the AN12 neuron. However, for the songbird model no quantitative fit to recorded data is available and model quality cannot be provided. On the other hand, the model displays similar features to the recordings in Fig. **??**.

For the model, we first calculated the spectrogram (Fig **??**a.2) of the target song (Fig. **??**a.1). As the recorded neuron is narrowband and responds best to low frequencies, we extracted the average power in the frequency band between $0-2$ kHz, denoted similarly to the AN12 example by $r_{exc}(t)$. The time course of an inhibitory neuron $r_{inh}(t)$ is computed as in the AN12 example by Equ. (**??**), depicted in Fig. (**??**a.4). The substraction of the inhibitory from the excitatory input gives rise to the firing rate of the model neuron $r_{model}(t) = r_{exc}(t) - r_{inh}(t)$ (Fig. **??**a.5). The relative amplitude of excitatory and inhibitory input is fixed to 1. Hence, the only free parameter is given by the time constant $\tau_{inh}$ of Equ. (**??**). We adjusted the time constant such that the response of the model neuron has a comparable time course to the recorded neuron (Fig. **??**c-e). A time constant of $\tau_{inh} = 40$ ms, the same one as chosen for the model of the auditory processing in the grasshopper, leads to an appropriate performance (Fig. **??**b).

The model neuron shows elevated firing rate (corresponding to additional spikes) in some periods when modulated noise is added to the target song. On the other hand, when broadband noise or chorus masker is added, firing rate is lowered for high noise amplitude. These results correspond to the observed response of the recorded neuron. Note that additional spikes in both recorded and model neuron occur preferentially shortly before a syllable onset. From the perspective of the model, this obversation can be explained as follows: A syllable with low-frequency content leads to subsequent suppression for $\sim 40$ ms. Input of similar amplitude as the preceding syllable can only be effective, i.e., eliciting additional spikes, after this period. Hence, additional spikes occur preferentially at the end of quietness periods. Similarly, subsequent suppression occurs preferentially in the second part of syllables. In contrast to the recorded neuron, the overall firing is elevated in the model neuron in response to the second half of the song motif. However, with respect to the simplicity of the model, overall performance is impressive. Similarly, taking the frequency band between $0-8$ kHz as input, the model shows comparable performance in reproducing the firing rate of a broadband neuron in response to the different masker conditions (not shown here).

[1]This part resulted from my student project in Woodshole 2006.

The fast-excitation-slow-inhibition model predicts particular properties of the neural system that can be tested.

- Songbirds could be trained to discriminate songs and time-scaled versions of the songs. If time-scaling brings the typical time-scale between syllables below 40 ms, discrimination capabilites should deteriorate.

- Noise could be added at specific positions within the song motifs. Noise snippets of, e.g., 10 ms should have different effects depending on their positions. If snippets would be positioned in front of some (but not all) syllables, the song could probably not be classified correctly anymore, as spike timing is shifted forwards for some syllables. However, noise snippets succeeding syllables should have no influence as they are suppressed by the preceding syllable. Note that in this case, the subsequent period of quietness should still be sufficiently sustained.

In the next section, we briefly discuss one example from the mammalian auditory system and then summarize the possible functions of the FexSin-circuit.

## 4.4 Sluggish response but precise firing?

How can single neurons integrate information on long time scales but yet maintain a rapid and precise response? This question is the so-called resolution-integration paradox (**?**). Recent studies try to explain this phenomenon (for a review see **?**).

The spectrotemporal receptive field (STRF) is a generalization of the spike triggered average (STA) including the spectral domain. Typical cells in the mammalian auditory cortex respond to a specific frequency band and simple temporal envelope with time constants of the order of 30 ms. An example from the mouse auditory cortex is given in Fig. **??**a. STRFs in field L of the songbird have very similar properties (**?**). However, the same kind of cells show very precise firing in response to stimulus onsets and other transients. In particular, when presented with both slow and fast modulations, auditory cells fire with a precision on the ms timescale (**??**b). Note that fast modulations alone are not sufficient to elicit precise firing: slow modulated envelopes gate the expression of fine structure (**?**).

A possible explanation of the integration-resolution paradoxon is given by Elhilali and colleagues (**?**). In phenomenological models they show that either synaptic depression or feedforward inhibition (with time constant $\tau = 65$ ms) preceded by fast excitation leads to comparable neural response in model

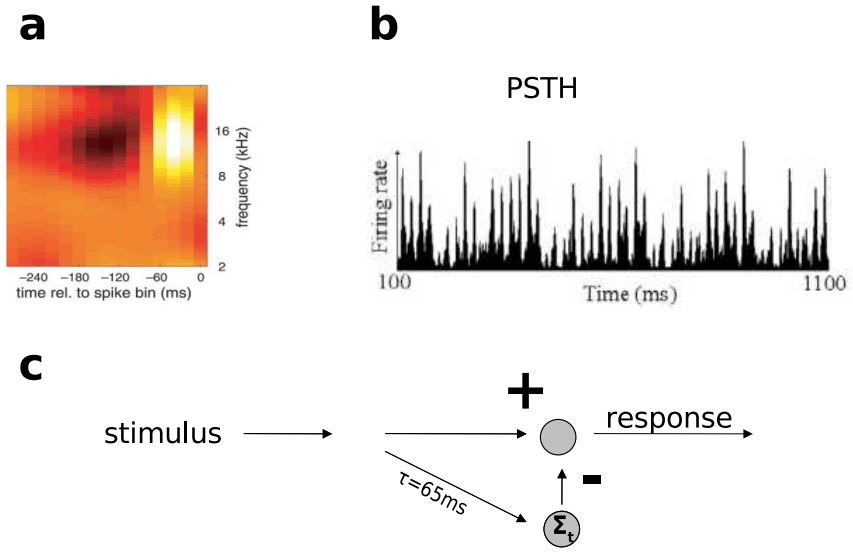


Figure 4.8: **Fast excitation and slow inhibition can explain precise firing in ferret cortex.** a) Spectrotemporal receptive field in mouse auditory cortex, adapted from **?**. b) Precise neural firing in the auditory cortex of ferrets, adapted from **?**. c) Scetch of model that is consistent both with a) and b).

neurons (Fig. **??**). Of course, the latter corresponds to the model circuit for the grasshopper and the songbird auditory system presented here.

## 4.5 Discussion: a canonical circuit for temporal processing?

We have seen that a seemingly simple circuit based on fast excitation and slow inhibition may the basis of a variety of phenomena in auditory processing:

- Fast excitiation and slow inhibition lead to a graded code (intraburst spike count) of a temporal feature (pause duration) that is a significant component of grasshopper communication signals.

- The circuit emphasizes temporal patterns in the presence of noise and can reproduce some properties of neural responses in the presence of

maskers in songbirds. Song motifs have similar statistics to speech, and hence, such a circuit may also be supportive in elucidating how humans can understand acoustic signals in presence of noise.

- Feedforward inhibition accounts for slow receptive fields but precise firing of cells in the auditory cortex of mammals.

Is such a circuit supported by anatomical and physiological evidence? In the grasshopper auditory system, anatomical studies suggest that the AN12 neuron receives both excitatory and inhibitory input (**?**), in particular the TN1 [2] and the UGN1 neurons are candidates for excitatory, the BGN1 and the UGN4 candidates for inhibitory input for the AN12. In field L of songbirds, local GABAergic interneurons driven by feedforward excitation from the thalamus seem to mediate delayed inhibition (**?**). In the auditory cortex of primates delayed inhibition (2-4 ms) has been measured with intracellular recordings (**??**). The inhibitiory conductance lasts up to 50-100 ms (**?**). We conclude that anatomical and physiological studies support the hypothesis of fast-excitation-slow-inhibition circuits.

This observed behavior of auditory neurons is only a necessary but not a sufficient condition to demonstrate the existence of a neural circuit with fast excitatory and slow inhibitory input. Synaptic depression and even intracellular adaptation may lead to similar results. However, at least in the AN12 neuron, firing-rate-dependent adaptation doesn't have a significant effect on the dynamics.

### Summary and Outlook:

What kind of circuit is responsible for the bursting response of the interneuron AN12? Here, we suggest a minimal circuit based on an interplay between fast excitation and slow inhibition (FexSin), thus marking the syllable onset but keeping the dependency on preceding pause duration. Electrophysiological and anatomical experiments are needed to authenticate this claim. We show that such a circuit can, in principle, also model neural response of auditory midbrain neurons under a variety of conditions. We discuss further results from the ferret cortex supporting the hypothesis that the FexSin circuit is an integral element of auditory systems.

[2] TN1 is known to be GABAergic (**?**)

# Chapter 5

# Song feature integration sufficient to account for behavioral response

We have studied a burst coding mechanism, modeled a plausible circuit for the AN12 neuron and hypothesized that the integration of the AN12 over a sufficiently long time is time-scale invariant. In this chapter, we try to relate this potential AN12 read-out mechanism to observed behavioral song classification.

## 5.1 Introduction

What syllable-pause combinations do grasshoppers prefer? With artificial model songs, one can test for appropriate syllable-pause combinations. An oval-shaped curve describes the syllable-pause combinations eliciting a behavioral response (Fig. **??**). Variation across animals is low for high syllable-pause ratios and high for low syllable-pause ratios.

Is the read-out of the AN12 sufficient to explain this behavioural response of female grasshoppers? The AN12 can explain the time-scale invariant output: If the sum of spikes is within the range of a certain target value, the behavioural response is positive, otherwise negative. However, the dependence on the ratio between syllable and pauses can be observed within a certain range only and stops beyond a certain period duration, i.e., syllable duration beyond 140 ms cannot elicit a response anymore (Fig **??**). Additionally, a classification based on AN12 integration would require a non-trivial post-AN12-synaptic computation: A simple thresholding would require a sufficient total spike count and therewith sufficient pause durations. How-

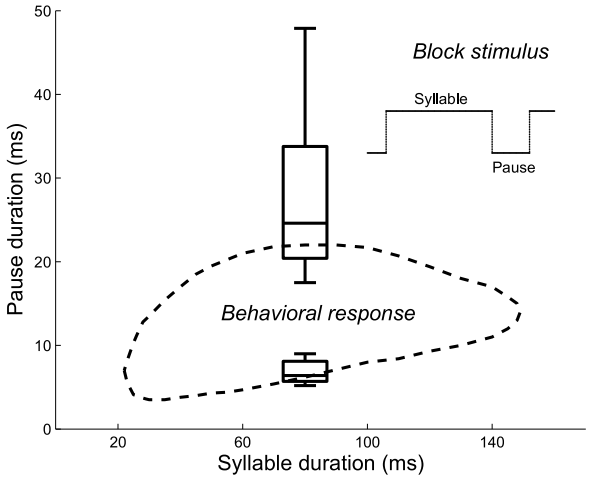


Figure 5.1: **Female *Chorthippus biguttulus* behavioural response.** The dashed curve depicts the area of syllable and pause values in artificial model songs to which female grasshoppers respond (one animal, 20% level). The boxes show the response variability across 17 females (20% level) at a given syllable duration of 80 ms. The variability is much lower for short pause durations (lower box) than for long pause durations (upper box).

ever, very high spike count implying very long pause duration should be suppressed. To account for this, additional feedback inhibition or similar mechanisms would be required. Instead, we suggest that the AN12 is complemented by other neurons constituting a feed-forward network where the read-out is based on integration and thresholding.

We put forward that the ascending neurons perform a parallel detection of relevant song features and that the head ganglion is evaluating the incoming spike trains to classify the stimulus. We construct a conceptual and numerical model in which 3 stimulus features are individually integrated over a fixed time window, thus calculating a moving average, and positively classified if the integration value crosses a certain threshold. If all 3 responses are positive, the overall motor response is permitted – corresponding to a logical AND operation.

## 5.2 Three thresholding operations

The 3 features to be integrated are overall pause duration, overall syllable duration and the syllable frequency (period count). The read-out neurons respond if the absolute integration passes a threshold value. Each threshold-

ing segments the syllable-pause space into permissible and non-permissible areas, effectively constituting a triangle (Fig **??**). As indicated by the arrows, only values inside the triangle lead to positive grashopper response. In detail, a) the overall integration of pause durations within $T_{dec}$ must have a sufficiently high value, allowing syllable-pause combination above the *pause-duration* line. Also, b) the overall integration of syllable duration within $T_{dec}$ must cross a certain threshold, corresponding to syllable-pause combination below the *syllable-duration* line. Furthermore, to limit the observed permissive values of absolute period durations, we introduced the integration of the number of syllable onsets, for clarity called period count. The period counting within $T_{dec}$ and thresholding allows only values to the left of the *period-count* line. Note that the slope of the *period count* line is fixed by requiring a constant sum of pause and syllable durations. All three features combined give a triangular range of admissable syllable-pause combinations that is related to the observations in behavioral experiments.

## 5.3 A minimal neural circuit for song classification

The model circuit is depicted in Fig **??**. Altogether, all information needed is transmitted via ascending neurons, whereas the integration and thresholding is performed at the subsequent postsynaptic read-out. As we have shown above, the pause integration is done by the summation of the AN12 spike train. For this part of the model, we keep the AN12 circuit with all parameters as described in chapter 4. For syllable integration, another ascending neuron is needed that responds tonically to syllables, i.e., proportional to syllable duration, and can then be summed up. Effectively, such a neuron simply counts syllable duration. The AN6 is a candidate neuron firing with constant firing rate in response to syllables (**?**). Furthermore, an ascending neuron that responds phasically to syllable onsets (with short time constant, e.g., $\tau_{phasic} = 3$ ms, and relative adaptation level at, e.g., 0.1 of the starting value) can be used as an input for a pulse counter. Here, the finite relative adaptation level is responsible for the smoothing the upper edge of the triangle range of admissable values: If syllable duration is substituted with pause duration, spike count will slightly decrease and lower the slope of that edge. No concrete evidence exists for the finite adaptation level of this phasic response neuron but – considering the large variance at the upper edge of the behavioral response curve (the upper box in Fig **??**) – the latter is no obligatory ingredient of pattern recognition. Note also that the spike train of the

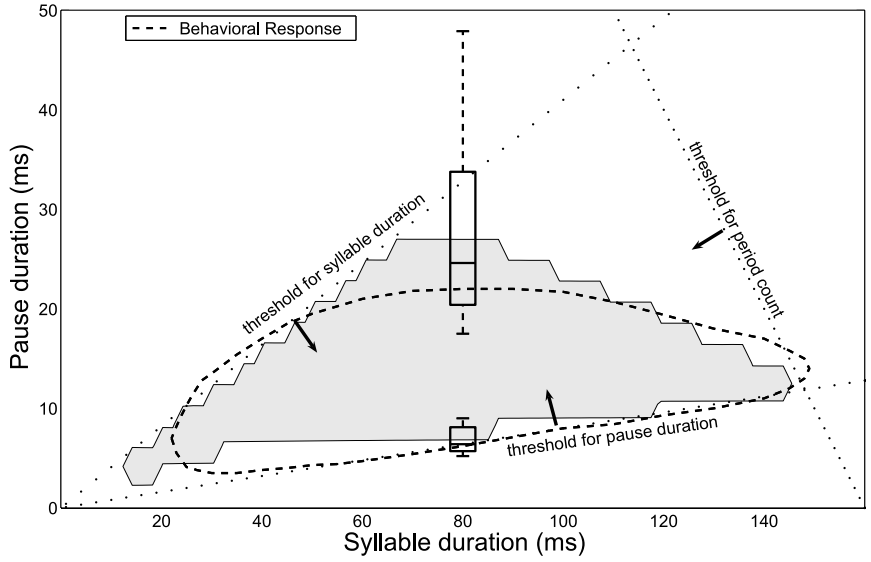


Figure 5.2: **The integration and thresholding hypothesis accounting for behavior of *Chorthippus biguttulus*.** Behavioral data is plotted as a reference as in **??**. A possible neuronal read-out mechanism would integrate over certain song features, i.e., total pause duration, total syllable duration and number of syllables and respond positively if all features cross threshold levels, constituting a triangle in the syllable-pause space. Threshold passing corresponds to getting inside the triangle as indicated by the arrows. The grey area depicts the read-out of a numerical model of ascending neurons (ANs). The AN12 measures pause duration, a tonically responding neuron measures syllable duration (AN6) and a strongly adapting neuron ($\tau$ = 3 ms) is responsible for the down-slope at high syllable durations. For details see Fig. **??**.

adapting ascending neuron can be seen as a combination of AN12 and AN6 spike trains. For example, consider that reading out the AN12 differently, taking bursts as a unit of information transmission (**?**), would be sufficient to constitute a period counter. However, as the pulse counter should be a distinct information channel, i.e., with distinct read-out, we choose to depict this information channel as a different neuron.

**Model specifications**. The AN12 neuron is specified as in chapter 4 with parameters unchanged. This is the only spiking neuron of the model. For the subsequent threshold operation, the output of AN12, termed $R_{AN12}$, is given

as the sum of all spikes within 1 s. The AN6 neuron is a firing rate based model, with firing rate $r_{AN6}(t) = 1\ s^{-1}$ when presented with a syllable and $r_{AN6}(t) = 0\ s^{-1}$ between syllables. For the subsequent threshold operation, the integrated output of AN6 is computed as $R_{AN6} = \int_0^{1s} r_{AN6}(t)dt$. The adapting neuron has firing rate $r_{phasic}(t)$ specified as a function of the stimulus $s(t)$ and an adaptation current $a(t)$.

$$\tau_{phasic}\frac{da(t)}{dt} = -a(t) + s(t)\ ,$$
$$r_{phasic}(t) = s(t) - A_{phasic}a(t)\ ,$$

where $\tau_{phasic} = 3$ ms and $A_{phasic} = 0.9$. For the subsequent threshold operation, $r_{phasic}(t)$ is also integrated over 1 s: $R_{phasic} = \int_0^{1s} r_{phasic}(t)dt$. All threshold operation are binary with threshold values

$$\theta_{AN12} = 8 \text{ spikes}\ ,$$
$$\theta_{AN6} = 0.72\ ,$$
$$\theta_{adapt} = 0.13\ .$$

The two last threshold values are dimensionless. $\theta_{AN6}$ can be interpreted as the minimum syllable-to-period ratio needed to elicit a response. Only if all three threshold are exceeded, i.e., $R_{AN12} > \theta_{AN12}$ and $R_{AN6} > \theta_{AN6}$ and $R_{phasic} > \theta_{phasic}$, the binary behavioral output is set to 1.

Modeling this system and chosing threshold values as specified above, one obtains a range of syllable-pause combinations similar to observed values (gray area in Fig. **??**). Our model is not sufficient to explain the lower left part of the syllable-pause space. In contrast to grasshoppers, our model predicts a positive behavioral response for very short pause and syllable durations. However, in this regime additional processes like gap detection (**?**) are operating, potentially dissolving this incoherence.

One prediction of the AN12 model is that higher syllable intensity leads to higher spike count. Hence, shorter time frames for evaluation ($T_{dec}$) should be sufficient to elicit a behavorial response in loud songs compared to quiet songs. This prediction is currently tested in behavioral experiments by the group of Bernhard Ronacher. Another prediction is displayed in Fig. **??**. If integration of ascending neurons is responsible for song recognition, then the detailed song structure is not important. However, it is only crucial that in average both syllable and pause durations are sufficiently long. Hence, a combination of two different mixtures of syllables and pauses – each on its own not sufficient to elicit behavioral response – should lead to a positive classification.

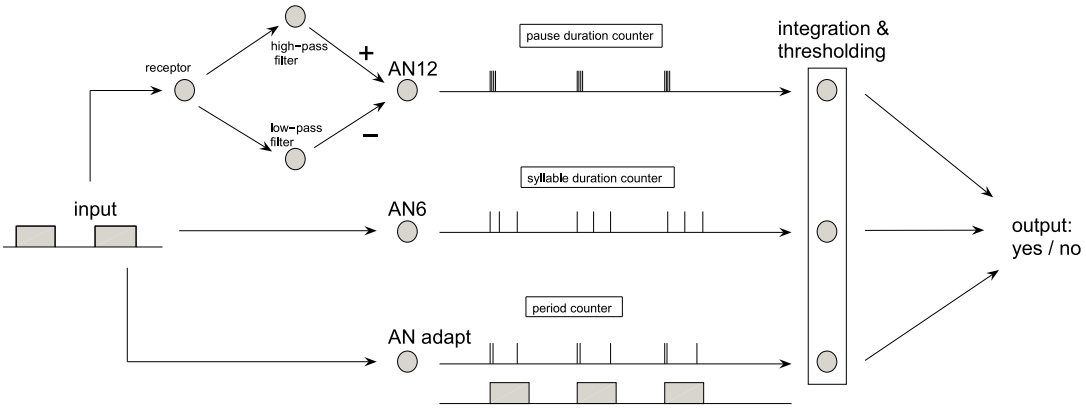


Figure 5.3: **Complete circuit accounting for the behavioral response.** A set of parallel feature detectors measures relevant sign stimuli which are then integrated and filtered by a logical AND-wiring. The AN12 measures total pause duration, the AN6 total syllable duration and the output of a hypothetical strongly adapting neuron is proportional to the number of syllables.

## 5.4 Discussion

A moving average of the AN12 response alone cannot explain all features of the behavioral response curve. We suggest a continuative model based on calculating a moving average from different sources. The whole behavioral response pattern could be explained by assuming integration from 3 different neurons within a certain time window $T_{dec}$, subsequent thresholding and a logical AND-operation which, again, could be implemented by integration and thresholding. In this model, one neuron accumulates absolute pause duration, another one absolute syllable duration and a third the number of periods (pulse counter) within $T_{dec}$. As we have shown, the AN12 encodes the individual pause duration by the spike count within a burst and the total pause duration by the total spike count. Another ascending neuron, the so-called AN6, fires tonically in response to ongoing syllables (**?**) and, hence, is the first candidate for syllable duration encoding. As a pulse counter, we propose a neuron with strong phasic response.

Our model corresponds well with some observations in behavioral experiments, as reported in **?**. Comparably to our study, the authors propose a conceptual model where one neuron detects syllable onsets, another neuron encodes syllable duration. Additionally however, in that study, the finite level of noise in the *pause* of natural songs was observed to play a role in behavioral response, indicating a more refined role of the syllable duration encoder. An

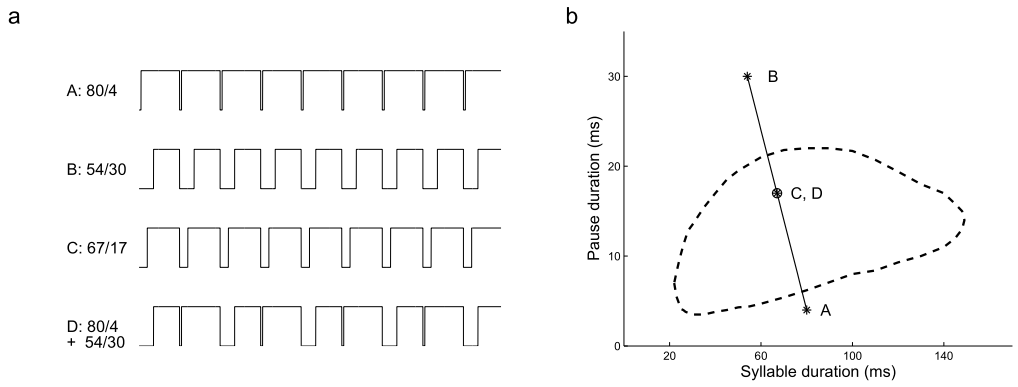


Figure 5.4: **Prediction of the integration hypothesis.** a) 4 different model songs. A and B would not elicit behavorial response as combinations of pause and syllable duration lie outside of the observed behavioral range. In contrast, model song C falls into the range. If integration over a time scale of, e.g., 750 ms is sufficient, only the average of syllable to pause ratios is decisive and model song D also elicits a response. b) The average syllable-to-pause combination of each model song.

extended numerical model of male mating song processing should take such effects but also gap detection into account. Possibly, amplitude fluctuations within songs on short time scales ($< 3$ ms), adaptation currents in receptor neurons and subsequent coincidence detection can adequately expand our model.

The integration and thresholding operation is consistent with an ample range of results in other systems. Some frogs communicate with a pattern of pulses and interpulse intervals. In the auditory midbrain of these frogs, a certain class of neurons integrates the number of acoustic pulses in a nonlinear fashion, responding only if a threshold number of appropriate pulses and interpulse intervals arrives (**??**). In contrast to the AN12 in the grasshopper, neurons in the anuran auditory midbrain are sensitive to the interpulse interval, i.e., the period, rather than to the preceding pause duration (**?**). In monkeys, the response of single neurons in the primary sensory cortex was correlated with behavioral output when discriminating vibrotactile stimuli (**?**). Only the integration of spike count over a time window with most mass within the first 250 ms correlated with behavioral output on a trial-to-trial basis. Furthermore, in electric fish, pacemaker neurons integrate incoming spike trains in a 'Sample-and-Hold' fashion over a wide range of time scales (up to several minutes) and adapt their firing rate in a graded manner, me-

diated by NMDA receptors and TRP channels (**?**). This demonstrates that plausible mechanism exist that can account for the hypothesized integration.

Our model of parallel feature-detection and subsequent thresholding is related to the multiple-look hypothesis (**?**) in which information from individual 'looks' is held in memory and combined later. According to our model, the information about individual looks of one specific feature is encoded in the associated ascending neuron and then combined by long time constant integration, similar to the pulse-integrator in (**?**).

### Summary and Outlook:

Here, we show that a seemingly complicated task like time-scale invariant pattern recognition can plausibly be mastered by the simple auditory system of grasshoppers by reducing the computation in a well ordered manner to parallel feature detection and subsequent integration. Behavioral data can be explained qualitatively as well quantitatively. A testable prediction is provided.

So far, the minimal model circuit for song classification is based only on recordings from a specific ascending neuron and general knowledge of other ascending neurons (**?**) and receptor cells (**?**). However, this circuit could be the basis for a canonical model of the grasshopper auditory system. Subsequently, experimental results would be incorporated and numerical studies could provide predictions that can, in turn, be tested in experiments. In the long run, such a model should explain

- the recognition of female songs, e.g., based on (**?**),
- gap detection – the SN6 and AN4 but also the AN12 could be involved in this task (**?**), and
- the interaction of high-frequency modulations and low-frequency envelope of communication signals: Is there a role for coincidence detection when reading out receptor cell response? Also in this context: what is the role of finite amplitude modulations on short time scales within pauses?

Furthermore, it will be a particular challenge to generalize the combination of the FexSin circuit and temporal feature integration to non-repetitive and complex temporal signals such as spoken words.

# Chapter 6

# Efficient coding: an introduction to information theory

What are the guiding principles of sensory processing? Here and in the following chapters, we will take a mathematical stance on this question. Information theory provides a suitable language to tackle issues related to coding (section 6.1). Representing an external signal, the neural system should use preferably short codes, not wasting ressources without need. This efficient coding hypothesis is treated mathematically by the source coding theorem (section 6.2). Furthermore, if information transmission is noisy, the code should be read-out in such a manner that reconstruction is as accurate as possible. This aspect can be formalized in the channel coding theorem (section 6.3). These hypotheses are not left unchallenged. For example, it is suggested that redundancy should be introduced to compensate for channel noise (**?**), seemingly in contradiction to efficient coding. However, such redundancy appears naturally when considering source and channel coding within one framework (section 6.4). Real neural systems are also confronted with highly complex signals while relying only on limited coding capacities. Hence, there is arguably need to obtain two objectives at the same time: limiting coding costs while maximizing reconstruction quality. This problem is formally solved by the rate distortion function (section 6.5). In contrast to these basic efficient coding approaches, it is argued that higher (cortical) representation can best be modeled by assuming a sparseness objective (**??**). Here, an overcomplete basis leads to a compact representation of natural signals by only a small number of spikes. Furthermore, our grasshopper studies emphasize the importance of feature detection (such as the pause duration of communication signal) and invariant decoding strategies (such as time-scale

invariant pattern recognition) in later stages of the auditory system. Hence, both the sparseness objective and the results from the grasshopper auditory system indicate that efficient coding per se is inadequate and, instead, that extraction of specific underlying causes may be crucial. Fortunately, the information bottleneck method, as an extension of the rate distortion function, introduces the notion of extracting relevant information (section 6.6), and can be regarded as the basis for integrating these slightly different theories. It is important to note that the mathematical theory of communiction (**?**) preceded and strongly influenced theories of neural coding in computational neurosciences (**??**). In fact, neurosciences were critized for relying solely on the information-theoretic description of neural systems and overemphasizing input-output relationships, considering the brain as a *computer gestalt* (**?**). We will shortly discuss the complementary but unexplored view of neural systems as autopoietic, selfregulating structures at the end of this thesis.

Most results of this chapter rely on (**???**). The information bottleneck method section is based on (**??**).

## 6.1 A measure for information

Given a set of $n$ possible events, each occuring with known probability $p_1, p_2, ..., p_n$. What is an appropriate measure to quantify the uncertainty of an event to occur? More precisely, we require that this measure, say $H$, obeys the following:

- Continuity: H should be continuous in $p_i$.
- Monotony: H should monotonically increase with the number of possible events $n$ if all occur with same probability. The intuition here is that more possibilities increase the uncertainty of outcome.
- Additivity: If the choice is broken down into successive choices the global measure of uncertainty $H$ should be the sum of the invidual values of $H$.

Shannon showed in his seminal work *A mathematical theory of communication* (**?**) besides many other important results that the only measure $H$ satisfying all three conditions barring a positive constant is the entropy:

$$H = -\sum_{i=1}^{n} p_i \log p_i \ .$$

Conveniently, the logarithm with base 2 is chosen. From hereon, log will always denote the logarithm with base 2.

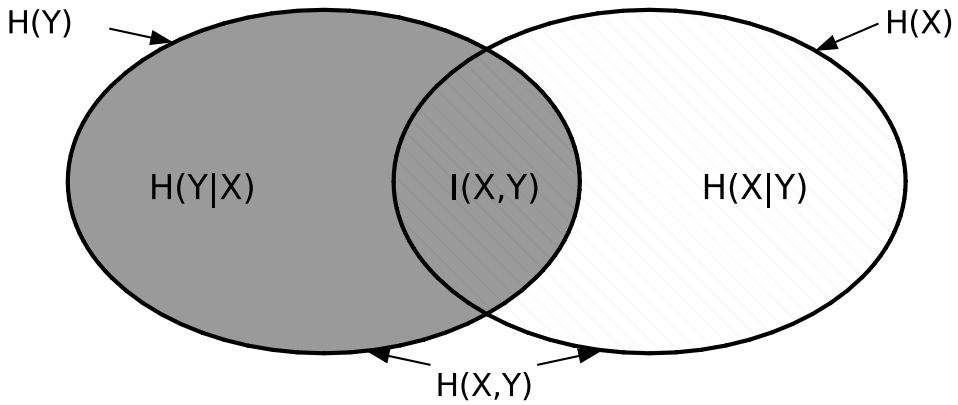


Figure 6.1: **Venn diagram illustrating the relationship between mutual information and entropic quantities.** Quite intuitively, the mutual information between two variables $X$ and $Y$ covers that area (information) that $H(X)$ has in common with $H(Y)$. Adapted from **?**. Note that the interpretation can become more difficult for three term entropies (**?**).

For the entropy it is true that $H(X) \geq 0$, where $X$ is a random variable. The entropy of a random variable is the average length of the shortest description of a random variable. Furthermore, if $X$ can take only discrete values, then $H(X)$ is the minimum expected number of binary questions required to determine the value of $X$. To summarize: the entropy is the average amount of information required to describe a random variable.

The concept can be extended to more than one random variable - then called joint entropy. Particularly interesting is the measure of how much information one random variable $X$ carries about another random variable $Y$. This measure is called mutual information and is formally given by

$$I(X,Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \ .$$

Mutual information is symmetric in $X$ and $Y$ and nonnegative. It can be written in terms of entropic quantities, e.g.,

$$\begin{aligned} I(X,Y) &= H(X) - H(X|Y) \\ I(X,Y) &= H(X) + H(Y) - H(X,Y) \ . \end{aligned}$$

These relations are illustrated in Fig **??**. Finally, we give the definition of the Kullback-Leibler distance, also called relative entropy:

$$D(p||q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)} \ . \tag{6.1}$$

We can see that the mutual information is a special case of the Kullback-Leibler distance:

$$I(X,Y) = D(p(x,y)||p(x)p(y)) \ .$$

## 6.2 Shannon source coding

We want to compress a random variable $X$ into a set of symbols, e.g., a binary code. This is the problem of data compression. What is the shortest code we can obtain on average?

To get a first intuition, let us look at a particularly simple example. Let $X$ be the color of marbles occuring with the following frequency and encoded by a binary word code:

$$\begin{aligned} P(X = red) &= \frac{1}{2}, \qquad C(red) = 0 \ , \\ P(X = blue) &= \frac{1}{4}, \qquad C(blue) = 10 \ , \\ P(X = orange) &= \frac{1}{8}, \qquad C(orange) = 110 \ , \\ P(X = green) &= \frac{1}{8}, \qquad C(green) = 111 \ . \end{aligned}$$

The entropy of $X$ is $H(X) = 1.75$ bits and the average word length $l$ of the binary code is also $E(l(X)) = 1.75$ bits. We see that the average length of a code can correspond to the entropy of the random variable. In fact, the entropy of $X$ is the shortest possible code (shortest expected word length) that can in principle be obtained such that the original source symbols can be exactly recovered (lossless source coding). More precisely:

**Theorem 6.2.1** *Source coding. Consider the Shannon code assignment* $l(x) = \log \frac{1}{p(x)}$. *Let* $L$ *be the associated expected length of the code* $L = \sum_x p(x)l(x)$. *Then*

$$H(X) \leq L \leq H(X) + 1 \ .$$

The proof is given on page 88 in (**?**). We should hasten to add that such a Shannon code assignment is not necessarily the optimal code, i.e., has the shortest average code word length. An optimal code is given by the so-called Huffman code (**?**). However, we can concatenate a given number $n$ of input symbols together $x^n = (x_1, x_2, ..., x_n)$ and use a joint codeword $l(x^n)$. The expected codeword length per input symbol is then $L_n = \frac{1}{n}E(l(X^n))$. Then we have:

**Theorem 6.2.2** *Source coding with block codes. The minimum expected codeword length per symbol satisfies*

$$\frac{H(X^n)}{n} \leq L_n \leq \frac{H(X^n)}{n} + \frac{1}{n} \ .$$

The proof is given on pages 88+89 in (**?**).

What happens if the expected description length is designed for the wrong distribution? Obviously, the code will comprehend some redundancy and the average expected word length will be larger than $H(X)$. Consider that the true probability distribution is $p(x)$. Our (wrong) estimation is $q(x)$. Then the minimal expected word length has a lower bound given by:

$$\begin{aligned} E(l(X)) &= \sum_x p(x) \log \frac{1}{q(x)} \\ &= \sum_x p(x) \log \frac{p(x)}{q(x)p(x)} \\ &= \sum_x p(x) \log \frac{p(x)}{q(x)} + \sum_x p(x) \log \frac{1}{p(x)} \\ &= D(p||q) + H(p) \ . \end{aligned}$$

Thus the redundancy in the code is equivalent to the Kullback-Leibler distance between true and estimated distribution.

A relevant example of a specific code is human language. If an alphabet has, e.g., 26 letters, this corresponds to a 26-dyadic code. When is it reasonable to introduce a new word into a given language? We may use source coding to obtain a heurestic ansatz for word introduction. An example: The English word *fairness* has found its way into the German language and is commonly used. Why has *fairness* been established? The German translation of *fairness* is *akzeptierte Gerechtigkeit und Angemessenheit*[1], a somehow complicated expression. The substitition of words has two effects: First, a long expression with relatively high total entropy is substituted by a shorter expression with lower entropy. Second, the introduction of *fairness* increases the total entropy of the language. Hence, the word-introduction prevails if the first effect is larger than the second effect[2].

[1]Source: http://de.wikipedia.org/wiki/Fairness, 08/2007, translation verified by Professor Jürgen Trabant, personal communication

[2]In fact, this example touches deep issues. A philosopher will ask whether concepts exist independent of language (the platonic view). More pragmatically, this *fairness*-example indicates the high redundancy of human language, also observed on other time scales. In fact, the example can motivate the use of an overcomplete (redundant) basis.

Comparably, one can ask: when should the brain supply additional coding space for a new concept? For example, the concept of *Claude Shannon* probably didn't have any particular representation (code word) in my cortex at the beginning of my studies but because of frequent use in relevant context this is different by now.

We are now in a position to introduce the notion of predictive coding. Assume that we observe the transmission of a sequence of events $x^n = (x_1, x_2, ..., x_n)$. Then we can encode $x_t$ relying on the information in $x^{t-1}$ thus saving coding space. A given joint distribution $p(x^n)$ is then efficiently encoded in a predictive sequential form

$$p(x^n) = \prod_{t=1}^{n} p(x_t|x^{t-1}) \ .$$

Indeed, the codelength of $x_1$ is $-\log p(x_1)$ and the codelength of $x_t$ is $-\log p(x_t|x^{t-1})$, resulting in a total codelength of

$$-\log p(x^n) = \sum_{t=1}^{n} -\log p(x_t|x^{t-1}) \ . \tag{6.2}$$

Crucially, we can achieve an optimal code of $x^n$ by predictive coding, i.e., an online-adaptive code (**?**) without relying on a block code. It is interesting to note that predictive information can - in some cases - be related to an explicit value function analytically. For example, in repeated horse races with non-adapting bookkeepers an optimally behaving gambler can get an increase in doubling rate (her pay-off) that is identical to the available predictive information (**?**).

## 6.3 Shannon channel coding

Up to now, we have only been concerned with efficient data encoding neglecting properties of real information channels. In fact, the above procedure is not much more than lossless compression. However, when studying biological systems we should include the notion of noisy channels.

A discrete channel is defined by an input alphabet, an output alphabet and a probability transition matrix $p(y|x)$, i.e., the probability of observing output $y$ given input $x$. Furthermore, we define the *information channel capacity* of a discrete memoryless channel as

$$C = \max_{p(x)} I(X, Y) \tag{6.3}$$

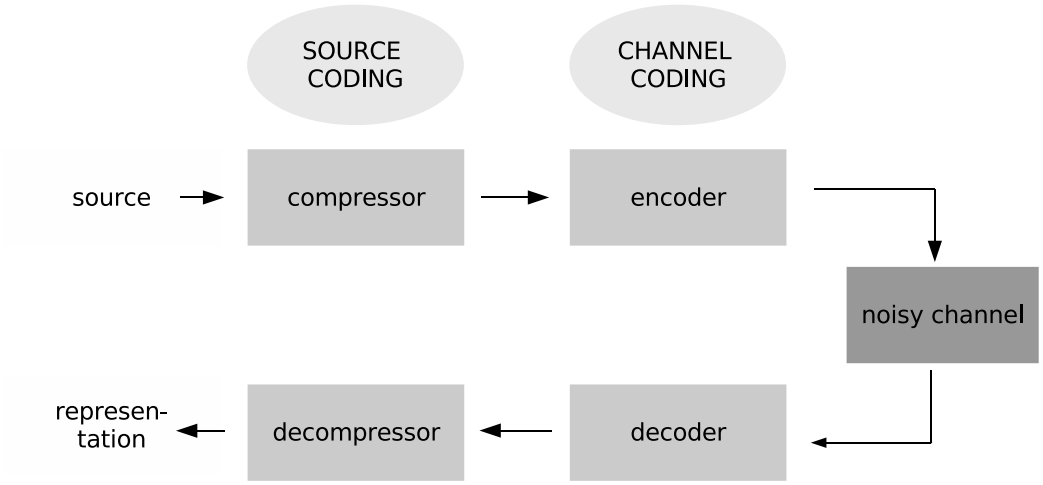


Figure 6.2: **Source and channel coding.** The original source is compressed and encoded, e.g., into a binary code. Source coding implicitly assumes that the communication channel is noisefree. In physical and biological systems, channels are noisy. However, error-free decoding is still possible up to a certain channel capacity. Adapted from **?**.

where the maximum is taken over all possible input distributions $p(x)$. When all messages are transmitted noisily one would assume that errorfree decoding is not possible. However, Shannon showed in his noisy-channel coding theorem that errorfree decoding can still be done (**?**): If information is transmitted at a rate (channel symbols per source message) below $C$, then it is always possible, to find a block code of size $N$ and a decoding algorithm such that the probability of incorrect decoding is arbitrarily small. Formally:

**Theorem 6.3.1** *Noisy channel coding. Take $S$ input symbols indexed by $i = (1, 2, ..., S)$ encoded by codewords $(X_1, X_2, ..., X_S)$. Each codeword has length $N$. The rate of this code is defined as $R = \frac{\log S}{N}$. Then all rates below channel capacity $C$ are achievable. For every rate $R < C$ and $\epsilon > 0$ there exists a block code of length $N$, with large enough $N$, such that the maximal probability of assigning a wrong output symbol $k$ to a given input symbol $i$, i.e. $k \neq i$, is lesser equal $\epsilon$.*

The intuition behind the channel coding theorem is that by increasing the block size, one increases the input and output alphabets. Then any particular input produces an output that is restricted to a small subspace of the output alphabet - the typical output given that input. Effectively, disjoint output sequences for a subset of inputs can be achieved. A detailed proof is given

in chapter 10 of **?**.

For illustration, take the noisy typewriter with 26 letters $\{A, B, C, ..., Z\}$ arranged in a circle. Each input letter is transmitted unchanged with probability $p = \frac{1}{2}$, or changed into the subsequent letter with $p = \frac{1}{2}$. Then the channel capacity is $\log 13$. Setting the block length $N$ to 1 and choosing to encode the subset of the alphabet $\{A, C, E, ..., Y\}$, this code has a rate of $\log 13$ and can be decoded without noise.

## 6.4 Source channel coding

So far, we have dealt with data compression and data transmission seperately. In fact, one can treat both problems sequentially as indicated in Fig **??**. However, is it possible to combine both problems? We have seen that source coding allows a compression $R > H$ and that channel coding leads to the requirement $R < C$. Is it necessary and sufficient to require that $H < C$ in order to send a source over a channel such that errorfree decoding is possible?

The joint source channel coding theorem states that both operations, data compression and data transmission, can be done in one stage. On the other hand, dealing with both problems separately is as efficient as considering both exercises together.

First take note of the Asymptotic Equipartition Property that can be regarded as the law of large numbers for information theory.

**Theorem 6.4.1** *The Asymptotic Equipartition Property states that if $X_1, X_2, ...$ are independent, identically distributed random variables $\sim p(x)$ then for every $\epsilon > 0$ there is an $N$ large enough such that a given sequence $x^N = (x_1, x_2, ..., x_N)$ belongs to a subset of all possible outcomes $A_\epsilon^N$ with probability $P > 1 - \epsilon$, $A_\epsilon^N$ has not more than $2^{NH(X)+\epsilon}$ members and the probability of each of its members $x_A^N$ is close to $2^{-NH(X)}$ in the following sense:*

$$2^{-N(H(X)+\epsilon)} \leq p(x_A^N) \leq 2^{-N(H(X)-\epsilon)} \ .$$

With this property we gain the notion of *typicality*, i.e., being a member of $A_\epsilon^N$. That has the advantage that we can deal with the limited subset of typical sequences, because the probability of non-typical sequences is arbitrarily low with $N$ large enough. We will make use of this property in the following theorem. Note that the Asymptotic Equipartition Property is equivalent to source coding (Theorem **??**).

**Theorem 6.4.2** *Source channel coding. Given a sequence of input symbols $X^N = (X_1, X_2, ..., X_N)$. The decoder estimates a sequence of output symbols as $\hat{X}^N = (\hat{X}_1, \hat{X}_2, ..., \hat{X}_N)$. The probability of decoding error is given by*

*$P^N_{error} = P(X^N \neq \hat{X}^N)$. If $X^N$ satisfies the Asymptotic Equipartition Property then there exists a source channel code with $P^N_{error} \to 0$ if $H(X^N) < C$. Otherwise, if $H(X^N) > C$ the probability of error has a lower bound $> 0$.*

**Proof.** We will encode only those sequences that belong to $A^N_\epsilon$. All other sequences appearing with probability $\epsilon$ will result in an error. Furthermore, there are at most $2^{N(H+\epsilon)}$ sequences in $A^N_\epsilon$ that can be encoded with at most $N(H+\epsilon)$ bits. The Asymptotic Equipartition Property also implies that for all $x^N \in A^N_\epsilon$ it is true that $-\frac{1}{N}\log p(x^N) \leq H(X) + \epsilon$. With this fact, we can choose the rate

$$R \equiv H(X) + \epsilon < C$$

such that probability of wrong decoding for a member of the typical set is $\leq \epsilon$. Hence,

$$\begin{aligned} P^N_{error} &= P(X^N \neq \hat{X}^N) \\ &\leq P(X^N \textit{ not} \in A^N_\epsilon) + P(\hat{X}^N \neq X^N \mid X^N \in A^N_\epsilon) \\ &\leq \epsilon + \epsilon = 2\epsilon \ . \end{aligned}$$

Joint source channel coding may have advantages for biological systems compared to compressing and transmitting data separately. For example, speech recognition degrades if the signal is transmitted in presence of white noise after redundancy reduction (compression). This suggest that redundancy in speech fits channel properties of the auditory system and the noisy environment. The grasshopper communication signal consists of more syllable-pause repetitions than minimally needed for song recognition. This additional redundancy may be appropriate to counteract the noisy environment (a channel property).

## 6.5 The tradeoff between effective communication and minimal coding cost

So far, we have treated discrete communication. However, biological signals are usually continuous in time and space. Continuity in the source signal poses new problems. For example, consider that the channel is noiseless. Then any real number can be transmitted with no error. Thus the channel has infinite capacity. Also, if the noise variance is non-zero and the input is continuous we can choose an infinite subset of inputs with arbitrary resolution such that the probability of error vanishes. But then, we also require infinite capacity. This is not plausible for a physical or biological channel. Hence, we

introduce a limitation on the capacity. A reasonable constraint, for example, is the maximum energy $S$ needed for any codeword $(x_1, x_2, ..., x_n)$:

$$s(x^n) \equiv \frac{1}{n} \sum_{i=1}^{n} x_i^2 \leq S \ .$$

The capacity is then a function of such an energy constraint. In general, for input $X$ and output $Y$ the capacity-cost function is

$$C(S) = \max_{p(x^n):E(s(x^n))\leq S} I(X,Y) \ . \tag{6.4}$$

This function is monotonically increasing, concave (decreasing positive slope), and for unconstrained optimization and finite input data equivalent to the capacity in **??**, here denoted as $C_{max}$. Its inverse function $S(C)$ exists for $0 \leq C \leq C_{max}$.

On the other hand, if we cannot transmit infinite information then the output can never reconstruct the input accurately – we encounter another challenge. The question then is: how can we achieve a *good* reconstruction. A criterium for *goodness* needs to be defined, called distortion. A distortion measure $d$ is an arbitrary distance measure between signal and reconstruction after signal transmission, i.e., a mapping from the combination of source and reconstruction alphabet into the set of non-negative numbers. Then we can ask: what is the minimal information rate needed to achieve a given distortion?

**Theorem 6.5.1** *Rate distortion function. For given independent identically distributed source $X$ with distribution $p(x)$, reconstruction $\hat{X}$ and distortion function $d(x, \hat{x})$ the minimal achievable rate at distortion $D$ is given by*

$$R(D) = \min_{p(\hat{x}|x):\sum_{x,\hat{x}} p(x)p(\hat{x}|x)d(x,\hat{x})\leq D} I(X,\hat{X}) \ . \tag{6.5}$$

For a proof see chapter 13 in **?**. The rate distortion function is monotonically decreasing and convex (decreasing value of negative slope) and if $D = 0$ equivalent to $H(X)$ from Shannon source coding.

The crucial observation is that the capacity-cost function (Equ. **??**) and the rate-distortion function (Equ. **??**) are related via the mutual information between input and output. This mutual information is the interface between data compression and data transmission in the general (continuous) case. The tradeoff between cost and distortion is then given by

$$S(D) = S(C = R(D)) \ .$$

The necessary and sufficient condition for optimality in data compression, i.e., removing source redundancy, and data transmission, i.e., minimizing reconstruction error, is consequently given by $C(S) = R(D)$. This is the generalization of the joint source channel coding theorem (**??**) and is called lossy source channel coding.

In rate distortion theory, we request – as a side constraint – that the average distortion is bounded: $E(d(x,\hat{x})) \equiv \sum_{x,t} p(x)p(\hat{x} \mid x)d(x,\hat{x}) \leq D$. Such a problem can be reformulated as a Lagrangian optimization by introducing a Lagrange multiplier. In this case, we have to minimize

$$\mathcal{L}(p(x,\hat{x})) = I(X,\hat{X}) + \beta E(d(x,\hat{x})) \ .$$

The stationarity conditions for this optimization are given by

$$\begin{aligned} p(\hat{x} \mid x) &= \frac{p(\hat{x})}{Z(\beta,x)} \exp\left(-\beta d(x,\hat{x})\right) \\ p(\hat{x}) &= \sum_x p(\hat{x} \mid x)p(x) \ . \end{aligned} \tag{6.6}$$

where the normalization function is given by $Z(\beta,x) = \sum_{\hat{x}} p(\hat{x}) \exp\left(-\beta d(x,\hat{x})\right)$. By iteratively updating both equations, the global optimum will be achieved. This algorithm is called the Blahut-Arimoto algorithm (**?**). The convexity of the problem guarantees optimal convergence as illustrated in Fig **??**a.

Lossy-source channel communication systems are optimal if the measured average distortion and average cost lie on the cost-distortion tradeoff curve. The average cost depends on the marginal distribution $p(x)$ at the channel input, the average distortion on the joint marginal distribution of source and reconstruction $p(s,\hat{s})$. The theorems above guarantee the possibility to find the correct marginal distribution by asymptotically long code words. However, in some cases source and channel already produce the correct marginal distributions - they are probabilistically matched. What are the requirements for probabilistically matching? Necessary conditions on optimal communication systems are given by the requirements on the cost function, which is dependent on the input distribution for the channel, $S(x)$, and the distortion function, which is dependent on the joint source-reconstruction distribution, $D(s,\hat{s})$ (**??**):

$$S(x) = D(p(y \mid x) \mid\mid p(y)) \tag{6.7}$$

$$D(s,\hat{s}) = -\log s \mid \hat{s} \tag{6.8}$$

up to shifts and scaling. Distributions fulfilling these conditions lie on the cost-distortion curve. Hence, also the long block codes of rate distortion

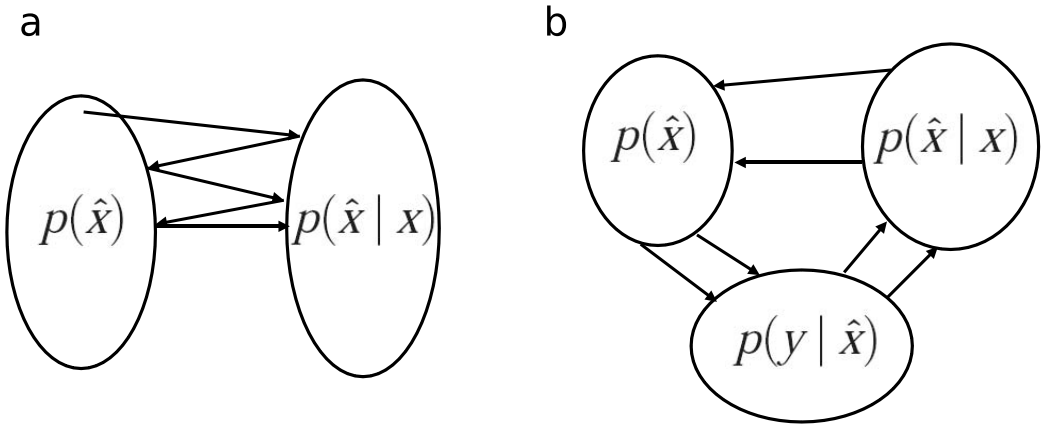


Figure 6.3: **Finding distances between convex sets.** a) The Blahut-Arimoto algorithm can be illustrated as finding the minimal distance (relative entropy) between two convex sets of probability distributions. This is achieved by iteratively finding the shortest distance between a point in one set and the opposite set. The resulting end point is taken as starting point for the next iteration. b) The information bottleneck algorithm iterates between three sets of probability distributions.

theory obey Equ. (**??**)+(**??**). However, for some source-channel pairs single letter codes without long block codes can be obtained with Equ. (**??**)+(**??**) leading to probabilistic source-channel matching (**?**).

## 6.6 The Information Bottleneck Method

The information bottleneck is a particular instance of rate distortion theory and a method for extracting relevant aspects of data (**?**). The latter aspect is particularly important for the description of biological systems, as filtering and interpretation of incoming data is arguably more crucial than straightforward representation. Technically, one seeks to capture those components of a random variable (input) $X$ that can explain observed values of another variable $Y$ (response). This task is achieved by compressing the variable $X$ into its compressed representation $\hat{X}$ while preserving as much information as possible about $Y$. The trade-off between these two targets is controlled by the trade-off parameter $\beta$. Hence, the information bottleneck problem can be formalized as minimizing the following Lagrangian:

$$\min \mathcal{L} : \mathcal{L} \equiv I(X, \hat{X}) - \beta I(Y, \hat{X}) \ .$$

The first term can be regarded as minimizing the complexity of the mapping while the second term tries to increase the accuracy of the representation. From the point of view of clustering, the information bottleneck method finds a quantization, or partition, of $X$ that preserves as much mutual information as possible about $Y$. From the perspective of machine learning, this corresponds to supervised learning. $X$ is the input signal and $Y$ tells you what aspects of the input should be learned. The information bottleneck method has been applied successfully in different circumstances, e.g., for document clustering (**?**), neural code analysis (**?**), gene expression analysis (**?**) and extraction of speech features (**?**).

It is clear that the information bottleneck method is a particular instance of rate distortion theory. The information bottleneck distortion measure is mutual information, i.e., Kullback-Leiber distance or information divergence (Equ. **??**). One can specify a class of rate-distortion problems where $X$ has to be compressed in such a way that information about another related variable $Y$ can be retrieved. Requiring a number of mathematical assumptions onto this class of rate distortion problems leads directly to the information bottleneck method (**?**).

Similarly to the rate distortion problem, one can also determine stationary conditions for the information bottleneck problem:

$$\begin{aligned} p(\hat{x}) &= \sum_x p(x)p(\hat{x} \mid x) \\ p(y \mid \hat{x}) &= \frac{1}{p(\hat{x})} \sum_x p(x,y)p(\hat{x} \mid x) \\ p(\hat{x} \mid x) &= \frac{p(\hat{x}}{Z(\beta, x)} \exp^{-\beta D(p(y|x)||p(y|\hat{x}))} \end{aligned} \tag{6.9}$$

By iterating over these 3 equations, one obtains an alternating optimizing projection algorithm, comparable to the Blahut-Arimoto algorithm (Equ. (**??**) and Fig. **??**b). However, in constrast to the Blahut-Arimoto algorithm, the information bottleneck algorithm can be trapped in local optima (**?**).

The information bottleneck has an additional important property: It works as a probabilistically matched source-channel. For this, note that the information bottleneck can be interpreted as follows: $Y$ is regarded as the target output distribution, and, in contrast to rate distortion theory, the information bottleneck already requires knowledge of input $X$ and output $Y$. Then the problem is to find the optimal representation $\hat{X}$ that relates input and output. In fact, the biological point of view assumes that evolution managed to achieve a probabilistic source-channel matching in biological communication systems – our task is then to find the internal characteris-

tics of this system. Observe, that by definition $I(X,Y) \geq I(X,\hat{X})$ and the Markov relation is given by

$$\hat{X} \leftrightarrow X \leftrightarrow Y \ .$$

In the following, we technically treat $\hat{X}$ as source and channel input. The effective information bottleneck distortion is given as

$$\begin{aligned} d_{IB}(x,\hat{x}) &= I(X,Y) - I(\hat{X},Y) \\ &= D(p(y \mid x) \mid\mid p(y \mid \hat{x})) \\ &\sim -\log p(\hat{x} \mid x) \end{aligned}$$

where we used Equ. (**??**) for the last step. The information cost can be demonstrated to be (**?**):

$$\begin{aligned} s(\hat{x}) &= D(p(x \mid \hat{x}) \mid\mid p(x)) \\ &= \beta D(p(y \mid x) \mid\mid p(y)) - \beta I(X,Y) - \log Z(\beta, x) \ . \end{aligned}$$

Thus, consistency with Equ. (**??**) and (**??**) is given and the information bottleneck finds an optimal representation $\hat{X}$ with respect to cost and distortion without block codes.

We finish this overview by introducing a particular realization of the information bottleneck problem: for Gaussian variables and linear mapping the problem can be analytically solved: the optimal functions are the solution of an eigenvalue problem (**?**). The key point is that the entropy of Gaussian variables can be written as the logarithm of the relevant covariance matrices between input and output. Minimizing the Lagrangian, finally, is equivalent to diagonalizing the covariance matrices; the eigenvector with the smallest respective eigenvalue gives the most informative part of the mapping between input and output[3]. This particular method will be used extensively in the next chapters. Relevant mathematical details will be introduced as required.

### Summary and Outlook:

Information theory provides the tools for solving data compression and data transmission problems. As neural systems presumably deal with both problems simultaneously, a joint source-channel coding approach should be appropriate. The information bottleneck emphasizes the role of a (hidden)

[3] An interesting question, raised in this context but not tackled within this thesis: How does the Gaussian information bottleneck relate to joint-source channel coding of the Gaussian channel with Gaussian input (**?**)?

internal state that is designed to optimally compress and transmit environmental signals. As motivated in the introduction, many real world problems are related to predicting future states of the environment. This assumption gives a natural interpretation of the target output $Y$ as the space of future events that can be predicted by past events. Furthermore, online adaptive coding is a natural implementation of a joint source-channel matching as seen in Equ. (**??**). In the following chapters, we will highlight this perspective and approach the past-future information bottleneck analytically.

# Chapter 7

# Local Predictive Coding and the Slowness Principle

Understanding the guiding principles of sensory coding strategies is a main goal in computational neuroscience. Among others, the principles of predictive coding and slowness appear to capture aspects of sensory processing. Predictive coding postulates that sensory systems are adapted to the structure of their input signals such that information about future inputs is encoded. Slow feature analysis (SFA) is a method for extracting slowly varying components from quickly varying input signals, thereby learning temporally invariant features. Here, we use the information bottleneck method to state an information-theoretic objective function for predictive coding. We then show that the linear case of SFA can be interpreted as a variant of predictive coding that maximizes the mutual information between the current output of the system and the input signal in the next time step. This demonstrates that the slowness principle and predictive coding are intimately related [1].

## 7.1 Introduction

One outstanding property of sensory systems is the identification of invariances. The visual system, for example, can reliably identify objects after changes in distance (**?**), translation (**?**), size and position (**?**). Neuronal correlates of invariance detection range from phase-shift invariance in complex cells in primary visual cortex (**?**) to high-level invariances related to face recognition (**?**). Hence, understanding the computational principles behind the identification of invariances is of considerable interest.

[1] This chapter is based on (**?**)

One approach for the self-organized formation of invariant representations is based on the observation that objects are unlikely to change or disappear completely from one moment to the next. Various paradigms for invariance learning have been proposed that exploit this observation (**?????**). As these paradigms extract the slowly varying components of sensory signals, we will refer to this approach as the slowness principle (**??**), in related literature also called temporal coherence or temporal stability principle (**???**). One formulation of this principle is Slow Feature Analysis (SFA; **?**). SFA has been successfully applied to the learning of various invariances in a model of the visual system (**?**) and reproduces a wide range of properties of complex cells in primary visual cortex (**?**). In combination with a sparseness objective, SFA can also be used as a model for the self-organized formation of place cells in the hippocampus (**?**; for related work see **?**).

A different approach to sensory processing is based on temporal prediction. For successful completion of many tasks our brain has to predict future states of the environment from current or previous knowledge (**?**). For example, when trying to catch a ball, it is not the current position of the ball that is relevant, but its position in the moment of the catch. We will refer to processing strategies that aim at performing this prediction as predictive coding. Predictive coding is the precondition for certain forms of redundancy reduction that have been applied successfully to model receptive fields in primary visual cortex (**?**) and surround inhibition in the retina (**?**). Redundancy reduction has been proposed as the backbone of efficient coding strategies and inherently relates to information theoretic concepts (**????**). However, to our knowledge an information theoretic framework for predictive coding has not yet been formulated.

In this work, we use the information bottleneck method (**?**), as introduced in the last chapter, to derive an information theoretic objective function for predictive coding. The information about previous input is compressed into a variable such that this variable keeps information about the subsequent input. We focus on Gaussian input signals and linear mapping. In this case, the optimization problem underlying the information bottleneck can be reduced to an eigenvalue problem (**?**). We show that the solution to this problem is equivalent to linear slow feature analysis, thereby providing a link between the learning principles of slowness and predictive coding.

## 7.2 Linear SFA

Slow Feature Analysis is based on the following learning task: Given a multidimensional input signal we want to find scalar input-output functions that

generate output signals that vary as slowly as possible but carry significant information. To ensure the latter we require the output signals to be uncorrelated and have unit variance. In mathematical terms, this can be stated as follows:

***Optimization problem 1****: Given a function space $\mathcal{F}$ and an $N$-dimensional input signal $X_t = [X_1(t), ..., X_N(t)]^T$ with $t$ indicating time, find a set of $J$ real-valued instantaneous functions $g_j(X)$ of the input such that the output signals $(Y_j)_t := g_j(X_t)$ minimize*

$$\Delta(Y_j) \equiv \langle \dot{Y}_j^2 \rangle_t \tag{7.1}$$

*under the constraints*

$$\langle Y_j \rangle_t = 0 \quad \textit{(zero mean)} \tag{7.2}$$

$$\langle Y_j^2 \rangle_t = 1 \quad \textit{(unit variance)} \tag{7.3}$$

$$\forall i < j : \langle Y_i Y_j \rangle_t = 0 \quad \textit{(decorrelation and order)} \tag{7.4}$$

*with $\langle \cdot \rangle_t$ and $\dot{Y}$ indicating temporal averaging and the derivative of $Y$, respectively.*

Equation (**??**) introduces the $\Delta$-value, which is a measure of the slowness of the signal $Y_t$. The constraints (**??**) and (**??**) avoid the trivial constant solution. Constraint (**??**) ensures that different functions $g_j$ code for different aspects of the input.

It is important to note that although the objective is the slowness of the output signal, the functions $g_j$ are instantaneous functions of the input, so that slowness cannot be enforced by low-pass filtering. Slow output signals can only be achieved if the input signal contains slowly varying features that can be extracted by the functions $g_j$.

If the function space $\mathcal{F}$ is finite-dimensional, the optimization problem can be reduced to a (generalized) eigenvalue problem (**??**). Here, we restrict $\mathcal{F}$ to the set of linear functions $Y_t = AX_t$, where $A$ is a $J \times N$-dimensional matrix. In the following, we also assume that input signals $X_t$ have zero mean. Then the optimal matrix obeys the generalized eigenvalue equation

$$A\Sigma_{\dot{X}} = \Lambda A \Sigma_X \,. \tag{7.5}$$

Here, $\Sigma_{\dot{X}} := \langle \dot{X}\dot{X}^T \rangle_t$ denotes the matrix of the second moments of the temporal derivative of the input signals and $\Sigma_X$ is the covariance matrix of the input signals. $\Lambda$ is a diagonal matrix that contains the eigenvalues $\lambda_j$ on the diagonal. The solution of the optimization problem for SFA is

given by the $J \times N$ matrix $A$ that contains the eigenvectors to the smallest eigenvalues $\lambda_j$ as determined by the generalized eigenvalue equation (**??**). For the mathematically interested reader, a derivation of equation (**??**) can be found in Appendix B.

We assume that the covariance matrix of the input data has full rank and is thus invertible. The generalized eigenvalue problem (**??**) can then be reduced to a standard left eigenvalue problem by multiplication with $\Sigma_X^{-1}$ from the right:

$$A \left[\Sigma_{\dot{X}} \Sigma_X^{-1}\right] = \Lambda A \,. \tag{7.6}$$

For discretized time the temporal derivative is replaced by $X_{t+1} - X_t$ and $\Sigma_{\dot{X}}$ can be rewritten as $\Sigma_{\dot{X}} = 2\Sigma_X - \left[\Sigma_{X_{t+1},X_t} + \Sigma_{X_t,X_{t+1}}\right]$, where $\Sigma_{X_{t+1},X_t} = \langle X_{t+1} X_t \rangle_t$ is the matrix containing the covariance of the input signals with the input signal delayed by one time step (**?**). Moreover, if the statistics of the input data is reversible, $\Sigma_{X_{t+1};X_t}$ is symmetric and $\Sigma_{X_{t+1};X_t} = \Sigma_{X_t;X_{t+1}}$. Using these relations in (**??**) yields

$$2A \left[ I - \underbrace{\Sigma_{X_{t+1};X_t} \Sigma_X^{-1}}_{=:\Sigma} \right] = \Lambda A \,. \tag{7.7}$$

Note that the eigenvectors of the SFA problem are also the eigenvectors of the matrix $\Sigma$ as defined in (**??**). Given the form of (**??**), we will be able to compare the eigenvalue problem with its counterpart from the information bottleneck ansatz of predictive coding.

## 7.3 Local predictive coding

The predictive coding hypothesis states that an organism extracts information from its sensory input that is predictive for the future (see e.g. **?**). Information-theoretically, this corresponds to mapping the data from the past into an internal state variable such that information between that state and the future data is maximized. To enforce a compact mapping, we introduce an additional penalty term that restricts the complexity of the mapping:

$$\max \mathcal{L} : \mathcal{L} \equiv I(state, future) - \beta^{-1} \; I(past, state) \;.$$

Obviously, the state variable cannot contain more information about the future than about the past, so that for $\beta^{-1} \geq 1$, the objective function $\mathcal{L}$ is negative: $\mathcal{L} \leq 0$. In this case, $\mathcal{L}$ is optimized by the trivial solution, where the state variable does not contain any information at all, because

then $\mathcal{L} = 0$. Thus, to obtain non-trivial solutions, the trade-off parameter should be chosen such that $0 < \beta^{-1} < 1$, or equivalently, $1 < \beta < \infty$.

The optimization problem above can also be formulated as an equivalent minimization problem that has the form of an information bottleneck as introduced in the previous chapter:

$$\min \mathcal{L} : \mathcal{L} \equiv I(past, state) - \beta \; I(state, future)$$

Here, we restrict ourselves to the special case of only one time step and a linear mapping. An extension to more time steps will be treated in chapter 9. Let us assume a discrete input signal $X_t$ that is mapped to an output signal $Y_t$ such that $Y_t$ is most predictive about the next input signal $X_{t+1}$ while minimizing the complexity in the information bottleneck sense, as illustrated in Fig **??**.

We assume that the input signal $X_t$ is an $n$-dimensional Gaussian vector and that the output signal $Y_t$ is generated by a noisy linear transformation

$$Y_t = AX_t + \xi \, .$$

The Gaussian white process noise $\xi$ is introduced for reasons of regularization: otherwise information theoretic quantities would diverge. For simplicity, we will assume that the noise is isotropic and normalized, i.e. that $\Sigma_\xi = \langle \xi \xi^T \rangle_t = I$, where $I$ denotes the unit matrix. This is no limitation, as it has been shown that every pair of $(A, \Sigma_\xi)$ can be mapped into another pair $(\hat{A}, I)$ such that the value of the target function $\mathcal{L}$ remains the same (**?**).

The above problem can now be stated in information-theoretic terms:

***Optimization problem 2****: Local predictive coding (LPC). Given input signal $X_t$ and output signal $Y_t = AX_t + \xi$ where $X_t$ and $\xi$ are Gaussian with $\langle \xi_t \xi_{t+1} \rangle_t = 0$, find the matrix $A(\beta)$ that minimizes*

$$\min \mathcal{L} : \mathcal{L}_{LPC} \equiv I(X_t, Y_t) - \beta \; I(Y_t, X_{t+1}) \; . \tag{7.8}$$

*with $\beta > 1$.*

The general solution to this problem has been derived in Chechik et al., 2005. (**?**). For completeness, a sketch of the derivation can be found in Appendix C. Here, we just state the solution:

**Theorem 7.3.1** ***Local Predictive Coding Information Bottleneck.*** *The solution to optimization problem 2 for Gaussian input signal $X$ with $Y =$*

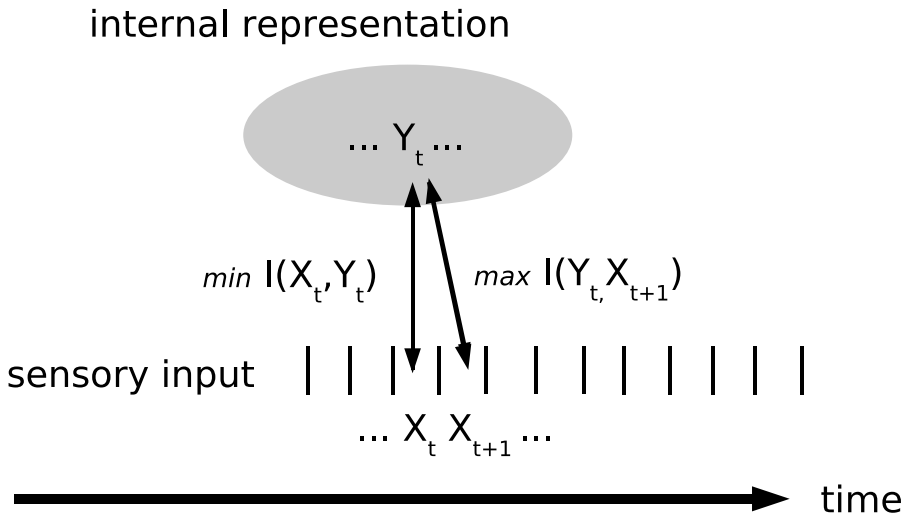


Figure 7.1: **Local predictive coding.** The sensory system compresses information of the current input $X_t$ into $Y_t$ such that the mutual information between $Y_t$ and the next input $X_{t+1}$ is maximized.

*$A(\beta)X + \xi$ is given by*

$$A(\beta) = \left\{ \begin{array}{cc} [0;\ldots;0] & 0 \leq \beta \leq \beta_1^c \\ [\alpha_1 W_1; 0; \ldots; 0] & \beta_1^c \leq \beta \leq \beta_2^c \\ [\alpha_1 W_1; \alpha_2 W_2; 0; \ldots; 0] & \beta_2^c \leq \beta \leq \beta_3^c \\ \vdots & \end{array} \right\}$$

*where $W_i$ and $\lambda_i$ (assume $\lambda_1 \leq \lambda_2 \leq \ldots$) are the left eigenvectors and eigenvalues of $\Sigma_{X_t|X_{t+1}}\Sigma_{X_t}^{-1}$, $\alpha_i$ are coefficients defined by $\alpha_i \equiv \sqrt{\frac{\beta(1-\lambda_i)-1}{\lambda_i r_i}}$, $r_i \equiv W_i \Sigma_X W_i^T$, 0 is an $m$ dimensional column vector of zeros, and semicolons separate columns in the matrix $A(\beta)$. The critical $\beta$-values are $\beta_i^c = \frac{1}{1-\lambda_i}$.*

The eigenvalues of $\Sigma_{X_t|X_{t+1}}\Sigma_{X_t}^{-1}$, $\alpha_i$ are guaranteed to be real and non-negative, as full-rank covariance matrices are positive definite. The key observation is that with increasing $\beta$ additional eigenvectors appear (second order phase transitions), corresponding to the detection of additional features of decreasing information content.

## 7.4 Relationship between slow feature analysis and local predictive coding

How does this solution relate to Slow Feature Analysis? We can rewrite $\Sigma = \Sigma_{X_t;X_{t+1}}\Sigma_{X_t}^{-1}$ in a more convenient form using Schur's formula:

$$\begin{aligned}\Sigma_{X_t|X_{t+1}}\Sigma_{X_t}^{-1} &= \left(\Sigma_{X_t} - \Sigma_{X_t;X_{t+1}}\Sigma_{X_t}^{-1}\Sigma_{X_{t+1};X_t}\right)\Sigma_{X_t}^{-1} \\ &= I - \Sigma_{X_t;X_{t+1}}\Sigma_{X_t}^{-1}\Sigma_{X_{t+1};X_t}\Sigma_{X_t}^{-1} \\ &= I - \left(\Sigma_{X_t;X_{t+1}}\Sigma_{X_t}^{-1}\right)^2 \\ &\overset{(??)}{=} I - \Sigma^2 \,,\end{aligned}$$

where we used the fact that time-delayed covariance matrices of reversible processes are symmetric. Note that the matrix $\Sigma = \Sigma_{X_t;X_{t+1}}\Sigma_{X_t}^{-1}$ also appears in the eigenvalue problem for linear SFA in the case of discrete time series (**??**), and hence, the optimal eigenvectors are the same for LPC and SFA. From (**??**) we know that the matrix to diagonalize in SFA is

$$\Sigma_{SFA} = 2I - 2\Sigma \tag{7.9}$$

with eigenvalues $\lambda_i^{SFA}$, whereas in LPC the target matrix is

$$\Sigma_{LPC} = I - \Sigma^2 \tag{7.10}$$

with eigenvalues $\lambda_i^{LPC}$. Solving (**??**) for $\Sigma$ and substituting the solution into (**??**), we obtain the relationship between the eigenvalues:

$$\lambda_i^{LPC} = \lambda_i^{SFA} - \frac{1}{4}(\lambda_i^{SFA})^2 \,.$$

SFA is guaranteed to find the slowest components first, whereas LPC finds the most predictive components first. For example, a very fast component can be very predictive, e.g. if the value at $t+1$ is the inverse of the current value (Fig **??**). Hence, from the local predictive coding point of view the absolute deviation from random fluctuations rather than slowness is important. This may be important for the analysis of discrete time series with high frequency components. However, this is only true for temporally discrete data: for continuous data one would expect a monotonous relation between eigenvalues of an information bottleneck approach and SFA eigenvalues.

Local predictive coding and SFA find the same components in the same order. The difference is that local predictive coding allows to quantify the components in terms of predictive coding. For example, take a 3-dimensional

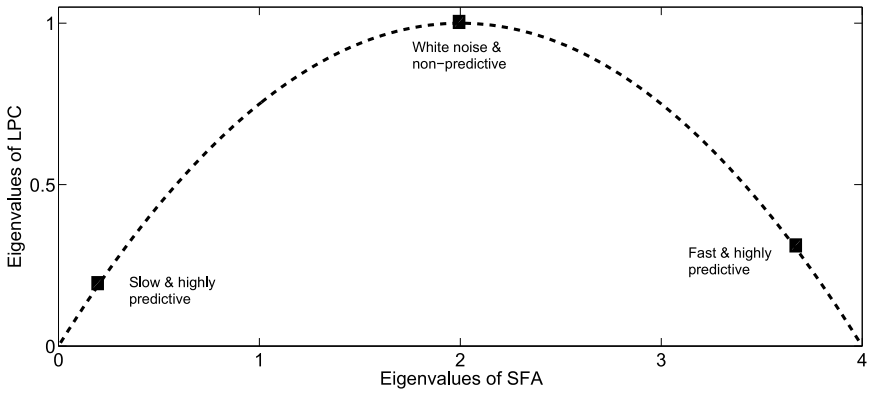


Figure 7.2: **Relationship between eigenvalues of slow feature analysis and local predictive coding.** For discrete time series fast components can be equally predictive as slow components. Only white noise is non-predictive.

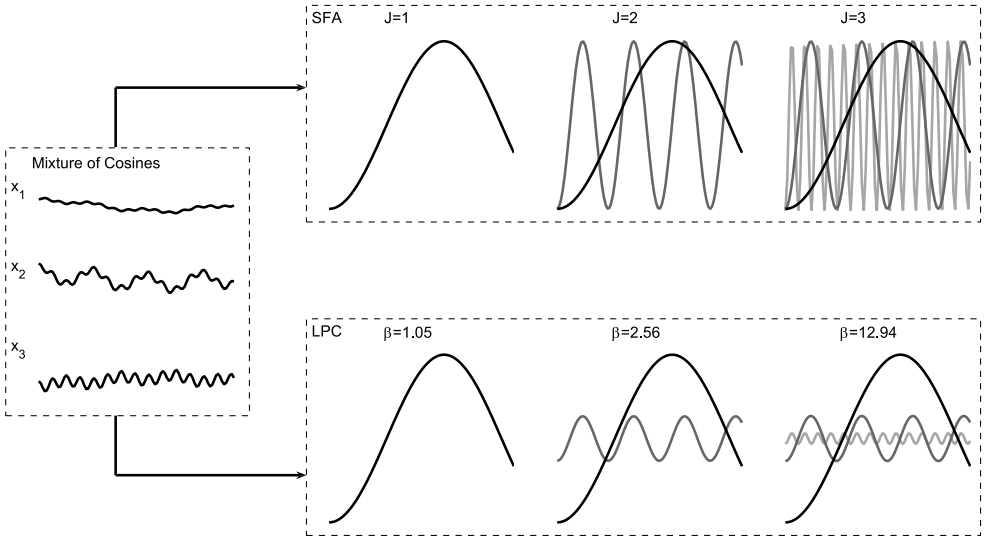


Figure 7.3: **Predictive coding and SFA differ in weighting of filtered components.** Both algorithms find the original cosines underlying a mixture signal. SFA discovers features in order of slowness only. Predictive coding give individual components weights in terms of their predictive information. For predictive coding, relative not absolute weightings are shown.

signal that consists of a mixture of cosines with different frequencies. Both methods can separate the original signals successfully (Fig. **??**). Slow feature analysis and local predictive coding reveal components in the same order, i.e., according to slowness. However, slow feature analysis accredits the same amplitude to all components while local predictive coding gives higher weights to slower components according to their predictive power.

## 7.5 Discussion

In this work, we relate slowness in signals to predictability. We have shown that predictive coding and slow feature analysis are equivalent for the restrictions of Gaussianity, linearity and one-time-step prediction. Both principles can explain some properties of visual receptive fields (**???**). On the one hand, our approach indicates that results from SFA studies such as the findings on complex cell properties (**?**) and hippocampal place cells (**?**) can be seen in terms of predictive coding. On the other hand, predictive coding by surround inhibition (**?**) and feedback connections (**?**) may be interpreted from the viewpoint of the slowness principle.

We have also shown that linear slow feature analysis can be motivated by information-theoretic principles. It is interesting to note that this linear, discrete case is also related to an implementation of second-order independent component analysis (**?**).

The relationship between predictive coding and temporal invariance learning has also been suggested in other work, e.g., by **?**, who argued that temporal invariance learning is equivalent to predictive coding if the input signals are generated from Ornstein-Uhlenbeck processes.

In one regard local predictive coding differs from slow feature analysis. The information bottleneck approach is continuous in terms of the trade-off parameter $\beta$ and new eigenvectors appear as second order phase transitions. The weighting of the eigenvectors is different in that it depends on their eigenvalue (Fig. **??**). This can be important when analyzing or modeling sensory systems where available band-width and hence, resulting signal-to-noise ratio is a limiting factor. For local predictive coding, available bandwidth, e.g., number of neurons, should be attributed according to relative amplitude, whereas slow feature analysis accredits the same bandwidth to all features.

We emphasize that our approach is not directly applicable to many real world problems. Our derivation is restricted to Gaussian variables and linear mappings. Both restrictions are not needed for SFA. Note that an extension of linear local predictive coding to non-Gaussian input signals would also capture the case of nonlinear processing, because after a nonlinear expan-

sion, the problem can be treated in a linear fashion. Usually nonlinear SFA corresponds to linear SFA after a nonlinear expansion of the input signals. In this sense nonlinear SFA can be regarded as the Gaussian approximation to the full non-Gaussian local predictive coding problem on the nonlinearly expanded input. This argument - together with effective nonlinear SFA models of the visual system (**??**) - suggests that sensory systems are tailored to extract (relevant) predictive information. For further research, we suggest to compare local predictive coding and slow feature analysis to generative hierarchical models for learning nonlinear statistical regularities (**??**).

The restriction on the immediate past implies that SFA does not maximize predictive information for other than first order Markovian processes. The generalization, i.e., relating the infinite past with the infinite future, can be best framed in terms of linear dynamical systems. Work on this topic is in preparation. Finally, predictive coding is not a stationary property of the evolved sensory system but dynamic and adapts with input statistics (**?**). A plausible extension of our work would aim to incorporate dynamic properties.

### Summary and Outlook:

Local predictive coding is defined as a tradeoff function where information about the next input is maximized such that information about the current input is minimized. We solve this problem for multivariate Gaussian signals and show that resulting eigenvectors are identical to those of linear slow feature analysis (SFA). Hence, we demonstrate that two seemingly different analytical frameworks have a joint basis. The predictive coding ansatz will be generalized in chapter 9.

# Chapter 8

# Sufficient system design

Engineers are concerned with the appropriate design of a system translating input into a target output. Living organisms, in contrast, need to learn and evaluate environmental statistics such that they can act and react to achieve their objectives such as surviving, reproducing or slaying the tedious fly. Engineers and biologists, hence, seem to deal with different concepts, but in fact both corresponding disciplines are significantly intersected. To see this, let us separate the organism's objective into two problems. The first is to extract the relevant statistics of the environment. The knowledge of these statistiscs is then the basis for the second problem: Interact with the environment such that external statistics are modified to the organism's advantage. As argued in the first chapter, predictive information can be interpreted as a low-level characterization of relevant information. That is, for the first problem mentioned above the organism has to learn the statistics of the incoming data stream such that generalization, e.g., extrapolating into the future data stream, is possible but overfitting is avoided. Learning, however, can be seen as finding a model that describes or even explains observations - with the usually unspoken assumption that the model will continue to be valid in the future (**?**). Furthermore, Rissanen pointed out that learning a model to describe a data stream can be interpreted as an encoding of those data (**?**). The number of bits required to encode the model parameters is called model complexitiy (**?**). This measure coincides with predictive information, the information the past of a data stream carries about the future of a data stream, as defined in **?**[1]. Indeed, predictive information can be regarded as a general measure for the complexity of a data stream, independent of parameterization (**?**). Altogether, in the framework of prediction, model estimation and encoding are equivalent.

[1]Note that Bayesian parameter estimation with a universal prior is also equivalent to the method of Rissanen (**?**).

The relationship between input and output data streams can generally be characterized by dynamical systems. Denote the input with $u(t)$ and the output by $y(t)$. A dynamical system in continous time is given by the functions $f$ and $g$:

$$\dot{x}(t) = f(u(t)) \tag{8.1}$$
$$y(t) = g(x(t)) = g(f(u(t))) \tag{8.2}$$

where $x(t)$ is called the state space. In the language of dynamical systems, the problem of parameter estimation is called *system identification*, i.e., identification of the functions $f$ and $g$. Dynamical systems are attractive because they provide a general description for continous and discrete data streams or time series. Crucially, we will see that the state space can be regarded as the bottleneck space summarizing the information that the past provides about the future.

Our primary concern here goes beyond the problem of system identification but also asks the question of sufficient system design. In general, coding a stochastic data stream – or equivalently estimating a model – with arbitrary precision requires infinite bits of information. Hence, we can hypothesize that the problem for a living organism in a stochastic and complex environment is to find an approximate but sufficient model of natural (spatio)-temporal statistics. *Complexity* of the environment can already be operationalized as the predictive information in the class of occuring natural signals (**?**), and, as we will see later, high complexity may be related to a high-dimensional or even infinite dimensional state space of the associated dynamical system. *Sufficiency* of a model means that those aspects are extracted that are relevant for the organism but other aspects are ignored. For the purposes of this PhD-study, we will focus on low-level signal extraction. By this we mean, that sufficient model design is not so much concerned with the content of the signal but only with predictability. Hence, the assumption is that a sensory system with limited bandwidth will focus on extracting those components of a signal that carry most information about the future. The advantage is that we don't have to specify ad hoc what is relevant for a specific animal. However, it is clear that predictability alone cannot explain sensory processing. In the grasshopper, for example, intraspecific signal patterns are reliably encoded whereas human speech, clinking across the field, probably has no accurate representation in the auditory system – although human speech carries a high amount of predictive information.

In this chapter, we introduce the concept of model reduction, i.e., the problem of finding a low-dimensional but somehow good approximation of the high-dimensional original model. The particular instances of balanced

model reduction and sub-space based identification methods will provide background for the subsequent chapter. There, we will focus on the question of how to reduce or design a system that uses the minimal amount of information between the past of a signal and an internal representation such that this internal representation carries sufficient information about the future of a signal.

The above mentioned second problem of the organism, manipulating the natural statistics to its own advantage, is not treated here. However, an extension of system theory, dynamical systems with control feedback loops, may constitute an appropriate extension that treats this second problem mathematically.

## 8.1 Linear systems

From hereon, we will focus on linear dynamical systems. In this section, we will define linear systems and introduce important properties. Some properties will be of use in chapter 9, other properties will help to understand the underlying concept. Proof of individual statements of this section can be found in (**?**).

A system is said to be a linear system, if any input-output pairs $(u_i(t), y_i(t))$ and $(u_j(t), y_j(t))$ can be combined such that $\alpha_i u_i(t) + \alpha_j u_j(t)$ leads to an output $\alpha_i y_i(t) + \alpha_j y_j(t)$ where $\alpha$ is a real constant. This is the so-called superposition property. If this property does not hold, the system is said to be nonlinear.

**Impulse response**. Any linear system with $p$ input variables and $q$ outputs variables can be described by

$$y(t) = \int_{t_0}^{t} G(t,\tau)u(\tau)d\tau \tag{8.3}$$

where the system is *relaxed* at $t_0$ when the output $y(t)$ is excited exlusively by input $u(t)$ with $t > t_0$. The system is also *causal* as the output $y(t)$ cannot be influenced by $u(\tau)$ with $\tau > t$. $G(t,\tau)$ is $q \times p$ matrix, called impulse response matrix, with entries $g_{ij}(t,\tau)$.

**State space**. Every linear system has a state-space description (Equ. **??**+**??**), specified for continous-time systems as:

$$\dot{x}(t) = Ax(t) + Bu(t) \ , \tag{8.4}$$

$$y(t) = Cx(t) + Du(t) \ . \tag{8.5}$$

and for discrete-time system as

$$x(t+1) = Ax(t) + Bu(t) \ , \tag{8.6}$$
$$y(t) = Cx(t) + Du(t) \ . \tag{8.7}$$

The state space $x(t)$ is $n$-dimensional, matrices $A$, $B$, $C$, $D$ must be $n \times n$, $n \times p$, $q \times n$ and $q \times p$ matrices, respectively. We only consider time-invariant systems, i.e., the matrices $A, B, C, D$ are not a function of time. This means that the same input leads to the same output, independent of when the input is applied.

**Transfer matrix**. Equation **??** is a convolution integral. It is possible to transform this convolution integral into a simple algebraic equation by applying the Laplace transform, $\overline{f}(s) = \int_0^\infty e^{-st} f(t)dt$, to **??**:

$$\overline{y}(s) = \overline{G}(s)\overline{u}(s) \ ,$$

where $\overline{y}(s)$ and $\overline{u}(s)$ denote the Laplace transforms of $y(t)$ and $u(t)$, and $\overline{G}(s)$ is the so-called transfer matrix of the system. The individual components $\overline{g}(s)$ of the matrix $\overline{G}(s)$ can be expressed in terms of poles $p_i$ and zeros $z_i$ in the so-called zero-pole-gain form:

$$\overline{g}(s) = k\frac{(s-z_1)....(s-z_m)}{(s-p_1)...(s-p_n)} \ . \tag{8.8}$$

where for poles: $|\overline{g}(p_i)| = \infty$.

Applying the Laplace transform to the state-space equations (**??**+**??**) yields:

$$\begin{aligned}\overline{x}(s) =& (sI - A)^{-1}B\overline{u}(s) \\ \overline{y}(s) =& C(sI - A)^{-1}B\overline{u}(s) + D\overline{u}(s) \ .\end{aligned}$$

These are algebraic instead of differential equations, i.e., differentiation and integration become multiplication and division, respectively. Hence, the transfer function can be related to the state-space description as

$$\begin{aligned}\overline{G}(s) &= C(sI - A)^{-1}B + D \\ &= \frac{1}{det(sI - A)}C[\text{Adj}(sI - A)]B + D. \end{aligned} \tag{8.9}$$

**Equivalence transformation**. The state space and system matrices are not unique. Consider $x' = Px$ where $P$ is an $n \times n$ real nonsingular matrix. Then the new state space equations

$$\begin{aligned}\dot{x}'(t) =& A'x'(t) + B'u(t) \\ y(t) =& C'x'(t) + D'u(t)\end{aligned}$$

with $A' = PAP^{-1},\ B' = PB,\ C' = CP^{-1},\ D' = D$ has the identical input-output relation as the original system (Equations **??** + **??**). The systems are called algebraically equivalent and $x' = Px$ is called an equivalence transformation.

**Bounded-input bounded-output stability**. An input $u(t)$ is said to be bounded if $u(t)$ does not diverge to positive or negative infinity. A system is called bounded-input bounded-output (BIBO) stable, if every bounded input results in a bounded output. A system with impulse response matrix $G(t)$ is BIBO stable, if and only if every $g_{ij}(t)$ is absolutely integrable in $[0, \infty)$:

$$\int_0^\infty |g_{ij}(t)|dt \leq \infty \ .$$

Equivalently, every pole of Equation (**??**) must have negative real part such that the transfer function is integrable. Using Equation (**??**), one can deduce that every pole of $\overline{G}(s)$ is an eigenvalue of $A$. If every eigenvalue of $A$ has a negative real part, then the associated continuous-time system is BIBO stable. However, not every eigenvalue of $A$ is a pole due to possible cancellation in Equation (**??**). A stronger condition is asymptotic stability: The equation $\dot{x} = Ax$ is called asymptotically stable if and only if all eigenvalues of $A$ have negative real parts. That can be seen by inspection the solution of $\dot{x} = Ax$: $x(t) = x_0 e^{At}$.

**Controllability.** Consider the state-space equation $\dot{x} = Ax + Bu$ with $n$ state space dimensions and $p$ inputs. Then the pair $(A, B)$ is said to be controllable if for any initial state $x_0$ and any final state $x_f$ there exists an input such that $x_0$ is transferred to $x_f$ in a finite time. Equivalently, the so-called $n \times np$ controllability matrix

$$[J_C]_n = \left( \ B \ AB \ ... \ A^{n-1}B \ \right)$$

must have full row rank $n$. Furthermore, if all eigenvalues of $A$ have negative real parts, the unique solution of

$$AW_C + W_C A^T = -BB^T$$

is positive definite. $W_C$ is the so-called controllability Graminan and can be calculated as

$$W_C = \int_0^\infty e^{A\tau} BB^T e^{A^T \tau} d\tau \ .$$

**Observability**. Consider the state space equations **??** + **??**. These equations are said to be observable if the unknown initial state $x_0$ can be determined by knowledge of $u(t)$ and $y(t)$ in $[0, t_f]$ with $t_f < \infty$. Equivalently, the observability matrix

$$[J_O]_n = \begin{pmatrix} C \\ CA \\ ... \\ CA^{n-1} \end{pmatrix}$$

has full column rank $n$. If all eigenvalues of $A$ have real negative part, the unique solution of

$$W_O = A^T W_O + W_O A = -C^T C$$

is positive definite with $W_O$ being the observability Gramian:

$$W_O = \int_0^\infty e^{A^T \tau} C^T C e^{A\tau} d\tau \ .$$

An intuitive interpretation and use of controllability and observability matrix will be given in the next section.

**Minimal realization**. A state space system description $(A, B, C, D)$ is said to be a minimal realization if $(A, B)$ is controllable and $(A, C)$ is observable. Hence, the state space dimension $dim(x)$ is as small as possible.

**Equivalence of continuous and discrete time systems**. It is important to realize that most results in continuous time are also valid in discrete time. Mathematically, this is due to the fact that functions analytic in the half-plane correspond to functions analytic in the unit disc (**?**). In our case, this means that asymptotic stability is guaranteed if all eigenvalues of $A$ have real negative part for the continuous case and if all eigenvalues of $A$ are inside the unit disc for the discrete case.

Note that we limit ourselves to linear time-invariant systems, thereby excluding many relevant non-linear or time-varying systems. However, the assumption of time-invariance is a good first approximation. Furthermore, nonlinear system can, under certain conditions, be linearized.

## 8.2 System identification

With our knowledge of linear systems we can now approach the concept of linear system identification. One can identify two classes of approaches

towards system identification (**?**). The first one is based on a Maximum Likelihood approach and was proposed by **?**. This approach is usually applied to autoregressive moving average models with exogeneous inputs (ARMAX). Using a prediction error framework, one can use the Least Square method (LS) to minimize the error between predicted and original output. This method is fully developed textbook knowledge (**?**).

Here, we will focus on the second class – in its advanced variants called subspace system identification methods. As subspace based methods are based on regression, they usually carry relatively low computational costs. Model order can be estimated directly. The method can be traced back to **?** and is developed in different variants as canonical variate analysis (CVA) in **?**, numerical methods for subspace state space identification (N4SID) in **?** and multivariate output error state space (MOESP) in **?**. Their key idea is based on the state space description of linear system or ARMAX models where the different variants give different weigthings to the state space structure (**?**). The state space summarizes all information of past input that is useful for mean square prediction. In the following, we introduce the main idea of subspace-based system identification by first presenting the main idea of **?** and then giving the general procedure

The problem of system identification is defined as finding the system matrices $(A, B, C, D)$ given input $u(t)$ and output $y(t)$. For this purpose consider the linear relation of a discrete-time system:

$$y_{fut} = H u_{past}$$

where the input past is given by $u_{past} = [u_0\ u_{-1}\ ...\ u_{-n}]$ and the output future by $y_{fut} = [y_0\ y_1\ ...\ y_n]$ and $n \to \infty$. $H$ is the so-called Hankel matrix, relating past and future of the system. The Hankel matrix has the following structure:

$$H = \begin{pmatrix} h_0 & h_1 & h_2 & ... & h_{n-1} \\ h_1 & h_2 & h_3 & ... & h_n \\ h_2 & h_3 & h_4 & ... & h_{n+1} \\ ... & & & & \end{pmatrix}$$

and the invidiual components can be estimated from the observed covariances: $h_k = \frac{1}{N} \sum_{t=0}^{N-k} y_{t+k} u_t^T$ and $N \to \infty$. A variety of methods exist to efficiently compute the Hankel matrix (**?**). We will encouter a particular method suitable for our purposes in the next paragraph. It is crucial to realize that the Hankel matrix is related to observability and controllability

matrix in the following way:

$$H = [J_O]_n[J_C]_n = \begin{pmatrix} CB & CAB & CA^2B & ... & CA^nB \\ CAB & CA^2B & CA^3B & ... & CA^{n+1}B \\ CA^2B & CA^3B & CA^4B & ... & CA^{n+2}B \\ ... & & & & \end{pmatrix}$$

as can be deduced from

$$\begin{aligned} x_0 &= [J_C]_n u_{past} \\ y_{fut} &= [J_O]_n x_0 \ . \end{aligned}$$

Equivalently to the Hankel matrix in discrete-time systems, a Hankel operator for continous time systems with finite rank can be defined. For both discrete and continous-time systems, the eigenvalues of the Hankel matrix or operator, respectively, can be computed from the product of Gramians (**??**). We will focus on another approach, computing the eigenvalues, called Hankel singular values, directly.

**Ho-Kalman algorithm.** How can the system matrices $(A, B, C, D)$ be inferred from $H$? The Ho-Kalman algorithm takes the following approach:

- Compute the SVD: $H = U\Sigma V^T$ where $\Sigma$ is identical to the Gramians in the so-called balanced realization, as we will see in the next section (Equ. **??** + **??**).
- Factorize: $H = U\Sigma V^T = U\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}V^T := [J_O]_n[J_C]_n$
- Solve for $A, B, C$, i.e.,

  $$B = [J_C]_1 \qquad C = [J_O]_1$$

  To compute $A$, define the submatrices $[J_C]_{1:n-1}$ and $[J_C]_{2:n}$ obtained from $[J_C]_n$ by deleting the last and first row respectively. Then

  $$A = [J_C]^+_{1:n-1}[J_C]_{2:n}$$

  where $()^+$ denotes the Moore-Penrose pseudoinverse.

**Subspace identification.** The main idea underlying the state-space approach is the fact that if the state space were known, the state space equations (**??** + **??**) could be used to determine the system matrices. The crucial problem then is to obtain a good estimate of the state space.

First, consider the estimation of the system matrices from the state-space description. Define the tail matrices $Y_t$, $X_t$ and $U_t$:

$$\begin{aligned} Y_t :=& [y_t, y_{t+1}, y_{t+2}, ...] \\ X_t :=& [x_t, x_{t+1}, x_{t+2}, ...] \\ U_t :=& [u_t, u_{t+1}, u_{t+2}, ...] \end{aligned}$$

Then every sample trajectory satisfies:

$$\begin{pmatrix} X_{t+1} \\ Y_t \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} X_t \\ U_t \end{pmatrix}$$

Then we use linear regression and solve by the Least Square method:

$$\min_{A,B,C,D} = \left\| \begin{pmatrix} X_{t+1} \\ Y_t \end{pmatrix} - \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} X_t \\ U_t \end{pmatrix} \right\|$$

obtaining the estimates

$$\begin{pmatrix} \hat{A} & \hat{B} \\ \hat{C} & \hat{D} \end{pmatrix} := \frac{1}{N} \begin{pmatrix} X_{t+1} \\ Y_t \end{pmatrix} \begin{pmatrix} X_t \\ U_t \end{pmatrix}^T \frac{1}{N} \left\{ \begin{pmatrix} X_t \\ U_t \end{pmatrix} \begin{pmatrix} X_t \\ U_t \end{pmatrix}^T \right\}^{-1}$$

with $N$ being the number of samples.

Canonical Correlation Analysis (CCA) was introduced by **?** and will be used for state space identification. Aim is to find a suitable basis for crosscorrelation between two random variables. Given $U$ and $V$, two zero-mean random variables of dimension $m$ and $n$. Find two special orthonormal bases $(u_1, ..., u_m)$ for $A$ and $(v_1, ..., v_n)$ for $B$ such that $E(u_i, v_j) = \rho_i \delta_{i,j}$ for $i, j \leq min(n, m)$. Requiring the $\rho_i$'s to be nonnegative and ordered in decreasing magnitude makes the choice of bases unique if all $\rho_i$'s are distinct. A specific implementation is given below.

But how can the state space be obtained from input and output data? In the following, we give a common variant of Canonical Correlation Analysis applied on the input past and output future, providing a so-called balanced state space.

- Define input past and output future at time $t$:

$$\begin{aligned} U_{past} =& [u_{-1}^T, u_{-2}^T, u_{-3}^T ...]^T \\ Y_{fut} =& [y_0^T, y_1^T, y_2^T, ...]^T \end{aligned}$$

- Normalize with Cholesky factors. Cholesky factors are given as $L_{past}L_{past}^T := \Sigma_u := \frac{1}{N}U_{past}U_{past}^T$ and $L_{fut}L_{fut}^T := \Sigma_y := \frac{1}{N}Y_{fut}Y_{fut}^T$. Normalized variables are then computed as: $\hat{U}_{past} := L_{past}^{-1}U_{past}$ and $\hat{Y}_{fut} := L_{fut}^{-1}Y_{fut}$.

- Perform SVD:
$$\frac{1}{N}\hat{Y}_{fut}\hat{U}_{past}^T =: \Sigma_{yu}\hat{V}\hat{\Sigma}\hat{W} \ .$$

- Compute the state space: $\hat{X}_t := \hat{V}^T\hat{U}_{past} = \hat{V}^T L_{past}^{-1}U_{past}$ and balance $\hat{X'}_t = \hat{\Sigma}^{\frac{1}{2}}\hat{X}_t$ such that $\frac{1}{N}\hat{X'}_t\hat{X'}_t^T = \hat{\Sigma}$.

The rows of $\hat{V}^T L_{past}^{-1}$ can be obtained directly as the left eigenvectors of $\Sigma_{yu}\Sigma_y^{-1}\Sigma_{uy}\Sigma_u^{-1}$ with $\rho^2$, the square of the canonical correlation coefficients, as eigenvalues (**?**). A formal proof of this relationship is provided in (**?**). Furthermore, note that the mutual information between input past and output future of stochastic time series is given as (**?**):

$$I(U_{past}, Y_{fut}) = \frac{1}{2}\sum_{i=1}^{n}\log\frac{1}{1-\rho_i^2} \ . \tag{8.10}$$

We will come back to this result in the next chapter.

## 8.3 Model reduction

If models are high-dimensional, engineers are often interested in suitable low-dimensional approximations to ease implementation. Similarly, organisms may also focus on extracting a low-dimensional representation of signal statistics. In system theory, this problem is called model reduction.

First, we introduce the general perspective (**?**). Consider the state space equations (**??** + **??**). Decompose the system matrices as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \qquad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} \qquad C = \begin{pmatrix} C_1 & C_2 \end{pmatrix} \ .$$

The truncated reduced order model is then defined as

$$\begin{aligned} \dot{x}_r(t) &= A_{11}x_r(t) + B_1u(t) \ , \\ y_r(t) &= C_1x_r(t) + Du(t) \ . \end{aligned}$$

With $A_{11}$ an $r \times r$ matrix, the order of the model is $r$. What is the quality of such a reduced model? One suitable measure is the additive error that is

given by the difference of the transfer functions:

$$\begin{aligned} G(s) - G_r(s) &= C(sI - A)^{-1}B + D - C_1(sI - A_{11})^{-1}B_1 - D \\ &= C'(s)\Delta^{-1}(s)B'(s) \ , \\ \Delta(s) &:= sI - A_{22} - A_{21}(sI - A_{11})^{-1}A_{12} \ , \\ B'(s) &:= A_{21}(sI - A_{11})^{-1}B_1 + B_2 \ , \\ C'(s) &:= C_1(sI - A_{11})^{-1}A_{12} + C_2 \ . \end{aligned}$$

This error is dependent upon the state coordinate basis of the system. Thus the art of model reduction is the identification of an appropriate basis. Recall that $(A, B, C)$ can be replaced with $(A' = PAP^{-1},\ B' = PB,\ C' = CP^{-1})$. Defining $P_L$ as the first $r$ rows of $P$ and $P_R$ as the first $r$ columns of $P^{-1}$, one obtains

$$A_r = P_L A P_R \ , \quad B_r = P_L B \ , \quad C_r = C P_R$$

Thus equivalence transformation, i.e., change of basis, and model reduction can be done in one step. In the following, we briefly discuss some important examples.

**Mode truncation.** A particular variant is called mode truncation. Here, one selects the transformation $P$ such that $A$ is diagonalized and selects the most dominant eigenvalues to keep the truncation error low (**?**).

**Hankel norm approximation.** Hankel norm approximation is important from an analytical point of view, as optimal approximations can always be achieved in this method. The Hankel norm is given as as

$$\|G(s)\|_H = \sup_{u \ \in \ L^2[0,\infty)} \frac{\|Hu\|_{u \ \in \ L^2[0,\infty)}}{\|u\|_{u \ \in \ L^2[0,\infty)}}$$

In his seminal work, **?** showed that it is always possible to find a reduced system such that $\|G(s) - G_r(s)\|_H$ is minimized.

**Balanced truncation.** For the purpose of this thesis, balanced truncation will be crucial. Let us obtain a particular realization, called balanced realization.

Consider the equivalence transformation $A' = PAP^{-1}, \quad B' = PB, \quad C' = CP^{-1}, \quad D' = D$ with $P$ a real nonsingular matrix. Then it is straightforward to show that the corresponding Gramians transform as

$$W_C' = P W_C P^T \qquad W_O' = (P^T)^{-1} W_O P^{-1} \ .$$

Every minimal realization can be transformed into a so-called balanced realization, i.e., controllability and observability Gramian are diagonal and equal. Such a transformation can be obtained as follows:

- Compute the singular value decomposition (SVD) of $W_O$: $W_O = U\Sigma U^T$ where $\Sigma$ is the diagonal matrix of singular values of $W_O$.
- Change basis with $P' = U\Sigma^{\frac{1}{2}}$. Hence,

$$W'_O = I \qquad W'_C = U\Sigma^{\frac{1}{2}} W_C \Sigma^{\frac{1}{2}} U^T \ .$$

- Compute the SVD $W'_C = V\Sigma^2 V^T$.
- Change basis with $P'' = \Sigma^{-\frac{1}{2}} V^T$. Then

$$W''_C = \Sigma^{-\frac{1}{2}} V^T V \Sigma^2 V^T \Sigma^{-\frac{1}{2}} = \Sigma \tag{8.11}$$

$$W''_O = \Sigma^{\frac{1}{2}} V^T I V \Sigma^{\frac{1}{2}} = \Sigma \ . \tag{8.12}$$

We can write

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \ ,$$

obtaining $\Sigma_1$ as controllability and observability Gramian after truncation at rank $r$ if the reduced order model is asymptotically stable. It is crucial to note that the basis for balanced realization is already obtained by the balancing step of system identification by canonical correlations analysis applied on past input and output future. Indeed, both subspace-based system identification by CCA and balanced model reduction ulitize Hankel singular values. Thus, including order selection into this CCA approach allows to combine the seemingly different issues of system identification and model reduction.

Finally, consider the error bound for balanced truncation. As shown by (**?**), the infinity norm of the absolute error is bounded as

$$\|G(s) - G_r(s)\|_\infty \leq 2 \sum_{k=r+1}^{n} \sigma_k \ .$$

Summary and Outlook:

In this chapter, we introduce linear dynamical systems theory. We explain some important concepts in order to give a framework of dynamical systems

theory and to provide a background for the next chapter. In particular, we describe the class of sub-space based system identification methods. For this, an efficient estimate of the state space can be obtained by canonical correlation analysis. The complexity of the model, i.e. here: the dimensionality of the state space, can be reduced by model truncation methods.

# Chapter 9

# The Past-Future Information Bottleneck of Dynamical Systems

Biological sensory systems need to encode and compress information simultaneously and in real time. As argued in the chapters 1 and 7, they should make use of temporal patterns of incoming signals such that the most predictive information is extracted. From an information-theoretic perspective, this can be regarded as joint lossy source-channel coding relying on adaptive predictive coding. Organism should do so such that sufficiently accurate prediction allows them to behave with resulting positive benefit while coding costs, e.g., energy ressources or requirement on the architecture of the neural system, are kept low. This is comparable to learning theory, where a complexity measure is desirable to quantify a preference for simpler models (**?**). Hence, the extraction of sufficiently accurate predictive information can also be regarded as the construction of an internal model mirroring external signal statistics but with limited complexity. Predictive information itself is a property of the observed data stream. This section's work aims not only to characterize predictive information in a signal, but to find *which properties* of the past are those that are relevant and sufficient for predicting the future. In particular, we describe the data stream as a dynamical system and seek to isolate the most predictive components of the past, relating them to parameters of the underlying system. In this framework, the concurrent tackling of system identification and model reduction is equivalent to joint lossy source channel coding in the temporal domain. In contrast to efficient sequential coding (Equ. **??**), only relevant (here: predictive) information but not stochastic fluctuations is encoded. Furthermore and unlike **?**, our approach allows an identification of predictive information independent of

observation time $T$.

The information bottleneck (IB) method, as introduced in chapter 6, is ideally suited to extract approximate minimal sufficient statistics (**?**). The two quantities of IB, namely compression level and relevant information, are complementary and in general we need to trade one for the other. The tradeoff between the two quantities is controlled by the $\beta$ parameter and makes apparent a natural order: By increasing $\beta$ one unravels features ("statistics") in $X$ that are informative about $Y$, where more informative features are revealed first.

Hence, the IB method is a natural approach to find the relevant past-future predictive features, defined above. In particular, given past signal values $U_p$ we are interested in compressing the information of the past into a model $\hat{Y}_f$ such that information about the future $I(\hat{Y}_f, Y_f)$ is preserved. When varying $\beta$ we obtain the optimal tradeoff curve – also known as the *information curve* – between compression and prediction, which is a more complete characterization of the complexity of the process. Our aim is to make the underlying predictive structure of the process explicit, and capture it by the states of a dynamical system. As a first step, we motivate our approach by extracting the predictive information of a time series. We will see that the state space is the natural bottleneck for predictive information. Hence, in the main part, we will aspire to compress a state space model of a dynamical system to maximize predictive information. We provide an analytic solution of the linear problem, on the basis of previously obtained results for the IB when the variables are jointly Gaussian (**?**). Our results show that as the tradeoff parameter $\beta$ increases, the compressed state space goes through a series of structural phase transitions, gradually increasing its dimension. Thus, for example, to obtain *little* information about the future, it turns out that one can use a one-dimensional (scalar) state space. As more information is required about the future, the dimension of the required state space increases up to its maximum $n$. The structure and location of the phase transitions turns out to be related to the eigenvalues of Hankel matrices which we have already encountered in the last chapter. Crucially, we will use a modified Ho-Kalman algorithm to obtain dynamical systems with information-theoretic optimally reduced state space. We also clarify the relation to canonical correlation analysis and characterize the optimal tradeoff function: the information curve. Finally, we characterize the information curve of the well-known spring-mass system, thus giving an example to demonstrate the numerical feasability of the past-future information bottleneck.

## 9.1 The state space as the natural bottleneck

In chapter 7, we have seen that extracting information about the subsequent time step can motivate the slowness principle. In this section, we generalize the approach to all past and future time steps. We want to find

$$\min \mathcal{L} : \mathcal{L} \equiv I(past, state) - \beta I(state, future) \ .$$

For this, define

$$U_p = \begin{pmatrix} u_t \\ u_{t-1} \\ ... \\ u_{t-(k-1)} \end{pmatrix} \ , \qquad U_f = \begin{pmatrix} u_{t+1} \\ u_{t+2} \\ ... \\ u_{t+k} \end{pmatrix} \tag{9.1}$$

with $u_t = [u_1(t), ..., u_p(t)]^T$ and $k \to \infty$. Then, we can formally state the following optimization problem.

***Optimization problem**: **Predictive coding of time series.** Given the signal $u_t$ with the past signal $U_p$ and the future signal $U_f$ as defined above and output signal $X_t = AU_p + \xi_t$ where $u_t$ and $\xi_t$ are Gaussian with $\langle \xi_t \xi_\tau \rangle_t = 0$ for $t \neq \tau$, find the matrix $A(\beta)$ that maximizes*

$$\min \mathcal{L} : \mathcal{L}_{PC} \equiv I(X_t; U_f) - \beta \ I(U_p; X_t) \ .$$

*with $\beta > 0$.*
This is equivalent to (**??**). The subsequent derivation, based on (**?**), holds for this case as well. We obtain:

**Theorem 9.1.1** ***Compression of the past into the state-space**. The solution to the optimization problem above for Gaussian input signal $u_t$ with $X_t = A(\beta)U_p + \xi_t$ is given by*

$$A(\beta) = \begin{Bmatrix} [0; \ldots; 0] & 0 \leq \beta \leq \beta_1^c \\ [\alpha_1 W_1; 0; \ldots; 0] & \beta_1^c \leq \beta \leq \beta_2^c \\ [\alpha_1 W_1; \alpha_2 W_2; 0; \ldots; 0] & \beta_2^c \leq \beta \leq \beta_3^c \\ \vdots & \end{Bmatrix} \tag{9.2}$$

*where $W_i$ and $\lambda_i$ (assume $\lambda_1 \leq \lambda_2 \leq \ldots$) are the left eigenvectors and eigenvalues of $\Sigma_{U_p|U_f} \Sigma_{U_p}^{-1}$, $\alpha_i$ are coefficients defined by $\alpha_i \equiv \sqrt{\frac{\beta(1-\lambda_i)-1}{\lambda_i r_i}}$, $r_i \equiv W_i \Sigma_{U_p} W_i^T$, $0$ is an $m$ dimensional column vector of zeros, and semicolons separate columns in the matrix $A(\beta)$. The critical $\beta$-values are $\beta_i^c = \frac{1}{1-\lambda_i}$.*

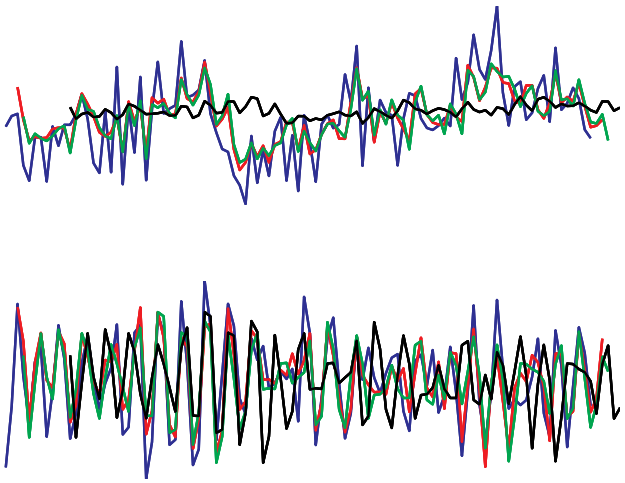

Figure 9.1: **Time series with dynamics of Eq (??).** Each panel represents the time course of one of the two components of vector $u_t$.Blue is the original time series, red the optimal 1-step ahead prediction, green the optimal 2-step ahead prediction and black the optimal 10-step-ahead prediction. The n-steps-ahead prediction utilizes only those states that date back n-steps or more.

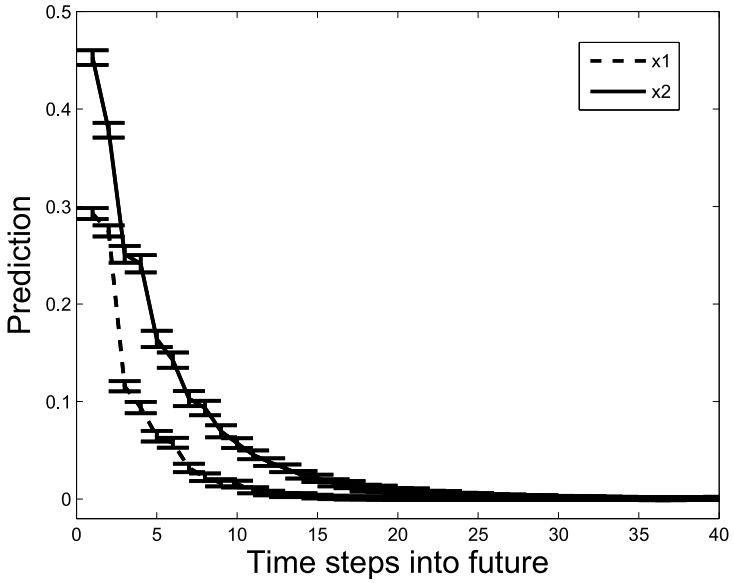


Figure 9.2: **Prediction with respect to time steps.** Prediction is defined as $\frac{a-<E>}{a}$ where $a$ is the average absolute amplitude of the timeseries $a = |< u_i(t)|>$ and $E$ the average absolute distance of the predicted time series to the original one.

To obtain an intuition of general predictive coding, we analyze a simple example. Assume $u_t$ to be a 2-dimensional signal $u_t = [u_1(t), u_2(t)]^T$ that can be written as a moving average model.

$$u_{t+1} = B_1 u_t + B_2 u_{t-1} + \xi_t$$

where $B_1$ and $B_2$ are 2 x 2 matrices, $\xi_t$ is a 2-dimensional vector with white noise $\xi_t \sim \mathcal{N}(0, 1)$. The concrete example has

$$B_1 = \begin{pmatrix} 0.1 & -0.2 \\ -0.5 & 0.2 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 0.5 & -0.3 \\ 0.1 & -0.9 \end{pmatrix}$$

$U_f$ depends only on $u_t$ and $u_{t-1}$, i.e., the relevant past is

$$U_p(t) = [u_1(t), u_2(t), u_1(t-1), u_2(t-1)]^T \ .$$

Hence, $A(\beta)$ should be 4-dimensional. Calculating $\Sigma_{U_p|U_f}\Sigma_{U_p}^{-1}$, we obtain 4 eigenvalues $< 1$. Then $X(t) = A(\beta)U_p(t)$ contains all information the past at time $t$ provides for all future time steps. In our case $X(t)$ is a 4-dimensional vector. One cannot read out directly the predicted value $\hat{u}(t+n)$ of future time steps $t+n$ from $X(t)$. However, one can extrapolate $\hat{u}(t+n)$ by established methods of parameter estimation such as the conditional least squares (CLS) estimation. The CLS estimate $P_n : \ \hat{u}(t+n) = P_n X(t)$ is obtained by $P_n = (X_t' U_{t+n}')(X_t' X_t)^{-1}$ where $X_t$ is the state space tracectory and $U_{t+n}$ the time series rescheduled $n$ steps ahead, similarly to Equ. (**??**). Some example predictions are depicted in Fig. (**??**). How good is the prediction after $n$ steps? Noise deteriorates the quality of prediction. In fact, predictive quality decreases exponentially with time steps (Fig. **??**). Note, that in contrast to the Kalman Filter, no updating algorithm is required. For an $n$ step prediction, one simply has to estimate $P_n$. Crucially, this naive IB ansatz on time series results in a compressed variable $X(t)$ that can be interpreted as the state space. This should suffice to motivate a direct information-theoretic treatment of dynamical systems. In fact, as we will see, this way of state-space identification corresponds to subspace-based system identification specified as canonical correlation analysis of the input past with the output future. We could identify the reduced systems based on theorem (**??**) and using least square estimation as introduced in chapter (8.2). However, here we chose a different approach and apply the information bottleneck directly on the Hankel matrix between input past and output future.

## 9.2 System reduction keeping predictive information

We turn to solving the past-future information bottleneck optimization problem of data streams:

$$\min \mathcal{L} : \mathcal{L}_{PFIB} \equiv I(past, model) - \beta I(model, future) \ . \tag{9.3}$$

in the general state space description. Again, we will rely on the Gaussian information bottleneck (**?**). We focus on the discrete-time case where the lumped linear dynamic system with process noise can be written as follows:

$$x_{t+1} = Ax_t + Bu_t \tag{9.4}$$
$$y_t = Cx_t + Du_t \tag{9.5}$$

Here $u$, $x$ and $y$ are $p \times 1$, $m \times 1$ and $q \times 1$ vectors and $A$, $B$, $C$, $D$ are $m \times m$, $m \times p$, $q \times m$, and $q \times p$ matrices, respectively. We denote the system parameters given by the above equations by $DS$. Our focus is on the bottleneck function of the state space, and, hence, we set $D = 0$, as it directly links the input to the output. Recall that the dimension $m$ of the state space corresponds to the number of poles of the transfer function (**?**). Since we are only interested in the effect of past input on future output at the present moment $t = 0$, we clamp the input to zero for times $t \geq 0$. Extensions, including also input future and output past into the analysis, are possible using the same techniques as, e.g., in **?**. However, as we gain only numerical accuracy in parameter estimation but no additional insight, we here stick to the direct input-output relation.

Assuming stationarity of the input signal, we can focus on the case where the past is measured up to $t = 0$, and the future for $t > 0$. Our aim is to find an optimal model output $\hat{y}_t$ that compresses the information of the input past but keeps information on the output future. The model output is specified as a function of $\beta$:

$$\hat{x}_{t+1} = A_r(\beta)\hat{x}_t + B_r(\beta)u_t \tag{9.6}$$
$$\hat{y}_t = C_r(\beta)\hat{x}_t + D_r(\beta)u_t + \boldsymbol{\xi} \ . \tag{9.7}$$

Hence, the IB Lagrangian can be written as

$$\min_{DS^r|DS} \mathcal{L} : \mathcal{L} \equiv I(U_p, \hat{Y}_f) - \beta I(Y_f, \hat{Y}_f) \tag{9.8}$$

where the input past, the output future and the model future are given by

$$U_p = \begin{pmatrix} u_t \\ u_{t-1} \\ ... \\ u_{t-(k-1)} \end{pmatrix} , \qquad Y_f = \begin{pmatrix} y_{t+1} \\ y_{t+2} \\ ... \\ y_{t+k} \end{pmatrix} , \qquad \hat{Y}_f = \begin{pmatrix} \hat{y}_{t+1} \\ \hat{y}_{t+2} \\ ... \\ \hat{y}_{t+k} \end{pmatrix}$$

with $u_t = [u_1(t), ..., u_p(t)]^T$, $y_t = [y_1(t), ..., y_q(t)]^T$, $\hat{y}_t = [\hat{y}_1(t), ..., \hat{y}_q(t)]^T$ and $k \to \infty$. The Lagrangian is optimized with respect to the matrices of the reduced system that are, in fact, a function of the tradeoff parameter $\beta$: $DS^r = (A_r(\beta), B_r(\beta), C_r(\beta))$. These will be derived in what follows.

We minimize Equ. (**??**). First, we can rewrite the mutual information quantities in terms of differential entropies.

$$\mathcal{L} = h(\hat{Y}_f) - h(\hat{Y}_f|U_p) - \beta h(\hat{Y}_f) + \beta h(\hat{Y}_f|Y_f) \tag{9.9}$$

For differential entropies, $h(X) = -\int_X f(x) \log f(x) dx$. In particular, for Gaussian variables

$$h(X) = \frac{1}{2} \log (2\pi e)^d |\Sigma_X|$$

where $|\Sigma_X|$ denotes the determinant of $\Sigma_X$ and $\Sigma_X := \langle XX^T \rangle_t$ is the covariance matrix of $X$ **?**. Hence, we have to find the covariance matrices of the quantities in Equ. (**??**). Recall that

$$x_0 = [J_C]_k U_p \ ,$$
$$Y_f = [J_O]_k x_0 \ ,$$

where the $m \times (p * k)$ controllability matrix and the $(q * k) \times m$ observability matrices are given, respectively, by

$$[J_C]_k = \begin{pmatrix} B & AB & ... & A^{k-1}B \end{pmatrix} , \qquad [J_O]_k = \begin{pmatrix} C \\ CA \\ ... \\ CA^{k-1} \end{pmatrix} .$$

Then we can calculate the Hankel operator

$$H = [J_O]_k [J_C]_k. = \begin{pmatrix} CB & CAB & CA^2B & ... & CA^kB \\ CAB & CA^2B & CA^3B & ... & CA^{k+1}B \\ CA^2B & CA^3B & CA^4B & ... & CA^{k+2}B \\ ... & & & & \end{pmatrix}$$

and thus $Y_f = HU_p$. Relying on Equ. (**??**+**??**) and adding model noise for regularization (**?**), we obtain $\hat{Y}_f = H(\beta)U_p + \xi$ for the model future. We seek

to identify $H(\beta)$, or equivalently, $A_r(\beta)$, $B_r(\beta)$, $C_r(\beta)$. Based on the Hankel operator, we compute $\Sigma_{\hat{Y_f}} = H(\beta)\Sigma_{U_p}H(\beta)^T + \Sigma_\xi$ and $\Sigma_{\hat{Y_f}|U_p} = \Sigma_\xi$. The last covariance matrix is given by:

$$\begin{aligned}\Sigma_{\hat{Y}_f|Y_f} &= \Sigma_{\hat{Y}_f} - \Sigma_{\hat{Y}_f,Y_f}\Sigma_{Y_f}^{-1}\Sigma_{Y_f,\hat{Y}_f} \\ &= H(\beta)\Sigma_{U_p}H(\beta)^T + \Sigma_\xi - H(\beta)\Sigma_{U_p,Y_f}\Sigma_{Y_f}\Sigma_{Y_f,U_p}H(\beta)^T \\ &= H(\beta)\Sigma_{U_p|Y_f}H(\beta)^T + \Sigma_\xi\end{aligned}$$

where we used Schur's formula in the first and last step (**?**). Neglecting irrelevant constants, Equ. (**??**) then becomes

$$\mathcal{L} = (1-\beta)\log|H(\beta)\Sigma_{U_p}H(\beta)^T + \Sigma_\xi| - \log|\Sigma_\xi| + \beta\log|H(\beta)\Sigma_{U_p|Y_f}H(\beta)^T + \Sigma_\xi|$$

Lemma A.1 in **?** states, that without loss of generality, we can set $\Sigma_\xi = I$. Then minimizing the Lagrangian gives

$$\frac{d\mathcal{L}}{dH(\beta)} = (1-\beta)(H(\beta)\Sigma_{U_p}H(\beta)^T)^{-1}2H(\beta)\Sigma_{U_p} + \beta(H(\beta)\Sigma_{U_p|Y_f}H(\beta)^T + I)^{-1}2H(\beta)\Sigma_{U_p|Y_f}$$

Equating this to zero and rearranging, we obtain conditions for the weigth matrix A.

$$\frac{\beta-1}{\beta}[(H(\beta)\Sigma_{U_p|Y_f}H(\beta)^T + I)(H(\beta)\Sigma_{U_p}H(\beta)^T + I)^{-1}]H(\beta) = H(\beta)[\Sigma_{U_p|Y_f}\Sigma_{U_p}^{-1}]\,. \tag{9.10}$$

Let us denote the singular value decomposion of the Hankel matrix as

$$H = W^T\Sigma_H V\ . \tag{9.11}$$

Then

**Theorem 9.2.1** ***The past-future information bottleneck of dynamical systems (PFIB).*** *The solution to Eq (**??**) is given by*

$$H(\beta) = W^T\Sigma_H(\beta)V\ , \tag{9.12}$$

*where* $\Sigma_H(\beta) = diag(\sigma_1(\beta), \sigma_2(\beta), ..., \sigma_m(\beta))$ *and* $\sigma_i(\beta) \equiv \sqrt{\frac{\sigma_i^2(\beta-1)-1}{r_i}}$ *and* $r_i = v_i\Sigma_{U_p}v_i^T$ *is the norm induced by* $\Sigma_{U_p}$ *and* $v_i$ *are row vectors of* $V$*.*

**Proof.** Let us first calculate $\Sigma_{U_p|Y_f}\Sigma_{U_p}^{-1}$, using Schur's formula:

$$\begin{aligned}
\Sigma_{U_p|Y_f}\Sigma_{U_p}^{-1} &= I - H^T(HH^T + I)^{-1}H \\
&= I - V^T\Sigma_H W(W^T\Sigma_H VV^T\Sigma_H W + I)^{-1}W^T\Sigma_H V \\
&= I - [V^T\Sigma_H^{-1}W(W^T\Sigma_H VV^T\Sigma_H W + I)W^T\Sigma_H^{-1}V]^{-1} \\
&= I - [I + V^T\Sigma_H^{-2}V]^{-1} \\
&= V^T\Sigma_H^{-2}V(I + V^T\Sigma_H^{-2}V)^{-1} \\
&= [(I + V^T\Sigma_H^{-2}V)V^T\Sigma_H^2V]^{-1} \\
&= (I + V^T\Sigma_H^2V)^{-1} \\
&= [V^T(I + \Sigma_H^2)V]^{-1} \\
&= V^T(I + \Sigma_H^2)^{-1}V \ . \qquad (9.13)
\end{aligned}$$

Hence, $\Sigma_{U_p|Y_f} = V^T(I + \Sigma_H^2)^{-1}V\Sigma_u$. Consider the positive definite bilinear form induced by $\Sigma_{U_p}$:

$$v_i\Sigma_{U_p}v_j^T = r_i\delta_{i,j}$$

where the $v_i$ are the row vectors of $V$. Denote $R$ as the matrix with $r_i$ on its diagonal. We substitute Equ. (**??** + **??** + **??**) into Equ. (**??**) and obtain

$$\begin{aligned}
&\frac{\beta-1}{\beta}[(W^T\Sigma_H(\beta)(I + \Sigma_H^2)^{-1}V\Sigma_{U_p}V^T\Sigma_H(\beta)W + I)(W^T\Sigma_H(\beta)V\Sigma_{U_p}V^T\Sigma_H(\beta)W + I)^{-1}]W^T\Sigma_H(\beta)V \\
&\qquad = W^T\Sigma_H(\beta)VV^T(I + \Sigma_H^2)^{-1}V\Sigma_{U_p}\Sigma_{U_p}^{-1}. \qquad \Leftrightarrow \\
&\frac{\beta-1}{\beta}[(W^T\Sigma_H(\beta)(I + \Sigma_H^2)^{-1}R\Sigma_H(\beta)W + I)(W^T\Sigma_H^2(\beta)RW + I)^{-1}]W^T\Sigma_H(\beta)V \\
&\qquad = W^T\Sigma_H(\beta)(I + \Sigma_H^2)^{-1}V.
\end{aligned}$$

By left-hand multiplication with $W$, inserting $W^TW$ between the brackets and right-hand multiplication with $V^T$, we obtain

$$\frac{\beta-1}{\beta}[(\Sigma_H(\beta)(I+\Sigma_H^2)^{-1}R\Sigma_H(\beta)+I)(\Sigma_H(\beta)^2R+I)^{-1}]\Sigma_H(\beta) = \Sigma_H(\beta)(I+\Sigma_H^2)^{-1} \ .$$

In this form, all matrices are diagonal and we can proceed in solving the individual Hankel singular values.

$$\frac{\beta-1}{\beta}\left(\frac{\sigma(\beta)_i^2r_i}{\sigma_i^2+1}+1\right)\left(\frac{1}{\sigma(\beta)_i^2r_i+1}\right) - \frac{1}{1+\sigma_i^2} = 0 \ .$$

After some reshaping, we obtain for $\sigma(\beta)_i^2$:

$$\sigma(\beta)_i^2 = \frac{\sigma_i^2(\beta-1)-1}{r_i} \qquad\qquad Q.E.D.$$

The reduced Hankel operator can be translated into reduced matrices $A(\beta)$, $B(\beta)$ and $C(\beta)$ by the algorithm of **?**. Define $\gamma(\beta)) \equiv [\Sigma_H(\beta)\Sigma_H^{-1}]^{\frac{1}{2}}$ and $[J_C(\beta)]_n = \gamma(\beta)[J_C]_n$, $[J_O(\beta)]_n = [J_O]_n\gamma(\beta)$. We can then factorize $H(\beta)$ into $H(\beta) = [J_O(\beta)]_n[J_C(\beta)]_n$. Then

$$B_r(\beta) = [J_C(\beta)]_1 \qquad C_r(\beta) = [J_O(\beta)]_1 \ . \tag{9.14}$$

Define the submatrices $[J_C(\beta)]_{1:n-1}$ and $[J_C(\beta)]_{2:n}$ obtained from $J_C$ by deleting the last and first row respectively. Then $A(\beta)$ can be computed as

$$A_r(\beta) = [J_C(\beta)]^+_{1:n-1}[J_C(\beta)]_{2:n} \tag{9.15}$$

where $()^+$ denotes the Moore-Penrose pseudoinverse. Similar to balanced model truncation (**??**), the PFIB procedure also relies on Hankel singular values. The difference is continous weighting in PFIB versus discrete weighting in balanced model truncation. Numerical evidence of the Ho-Kalman construction is provided in section 9.4.

## 9.3 Relation to CCA

In canonical correlation analysis, the eigenvectors and eigenvalues of $\Sigma_{Y_fU_p}\Sigma_{Y_f}^{-1}\Sigma_{U_pY_f}\Sigma_{U_p}^{-1}$ are computed. In the past-future information bottleneck the target matrix is $\Sigma_{U_p|Y_f}\Sigma_{U_p}^{-1} = I - \Sigma_{Y_fU_p}\Sigma_{Y_f}^{-1}\Sigma_{U_pY_f}\Sigma_{U_p}^{-1}$. Hence, eigenvectors of both procedures are identical (**?**). The eigenvalues of CCA, called (squared) canonical correlation coefficients, are denoted as $\lambda_i^{CCA} = \rho_i^2$. The eigenvalues of $\Sigma_{U_p|Y_f}\Sigma_{U_p}^{-1}$ can be calculated from Equ. (**??**) as $\lambda_i^{PFIB} = (1+\sigma_i)^{-1}$. Hence, the relationship between the Hankel singular values and the canonical correlation coefficients can be calculated by $(1+\sigma_i)^{-1} = \lambda_i^{PFIB} = 1 - \lambda_i^{CCA} = 1 - \rho_i^2$ to give

$$\rho_i^2 = \frac{\sigma_i^2}{\sigma_i^2+1} \qquad \text{or} \qquad \sigma_i^2 = \frac{\rho_i^2}{1-\rho_i^2} \ .$$

## 9.4 Information curve of predictive information

The information curve illustrates the tradeoff between model accuracy, here: predictive information $I(\hat{Y}_f, Y_f)$, and model complexity, here: required or compressed information from the past $I(U_p, \hat{Y}_f)$. This curve is similar to the rate-distortion curve of lossy source coding (chapter 6.5). As can be deduced

from **?**, the theoretical information curve for PFIB is given as

$$I(U_p, \hat{Y}_f)_\beta = \frac{1}{2}\sum_{i=1}^{n(\beta)} \log\left(\beta - 1\right)\sigma_i^2 \ , \tag{9.16}$$

$$\begin{aligned} I(\hat{Y}_f, Y_f)_\beta &= I(U_p, \hat{Y}_{fut}) - \frac{1}{2}\sum_{i=1}^{n(\beta)} \log \beta \frac{\sigma_i^2}{\sigma_i^2 + 1} \\ &= \frac{1}{2}\sum_{i=1}^{n(\beta)} \log \frac{\beta - 1}{\beta}(\sigma_i^2 + 1) \\ &= \frac{1}{2}\sum_{i=1}^{n(\beta)} \log \frac{\beta - 1}{\beta}\frac{1}{1 - \rho_i^2} \end{aligned}$$

where $n(\beta)$ indicates the maximal index $i$ such that $\beta \geq 1 + \frac{1}{\sigma_i^2}$.

As $\beta \to \infty$ the predictive information converges to

$$I(\hat{Y}_f, Y_f)_\infty = \frac{1}{2}\sum_{i=1}^{n} \log \frac{1}{1 - \rho_i^2}$$

Assuringly, this is identical to the Akaike result on the mutual information between past and future of a stochastic system (Equ. **??**, **?**).

In the next section, we will investigate one particular system and show numerically that the reduced systems given by equations (**??**+**??**) lie indeed on the information curve.

## 9.5 The spring-mass system

As an example system we apply the past-future information bottleneck to a spring-mass system with two different masses, both fixed with a spring $k_1$ at the wall, a spring $k_2$ connects the two masses. Two forces $u_1$ and $u_2$ perturb the masses such that they are displaced by $y_1$ and $y_2$ from their idle position (Fig. **??**). This can be modeled as a dynamic system with $A, B, C$ as

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{-(k_1+k_2)}{m_1} & \frac{-c}{m_1} & \frac{k_2}{m_1} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{k_2}{m_2} & 0 & \frac{-(k_1+k_2)}{m_1} & \frac{-c}{m_2} \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 \\ \frac{1}{m_1} & 0 \\ 0 & 0 \\ 0 & \frac{1}{m_2} \end{pmatrix}, \tag{9.17}$$

$$C = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{9.18}$$

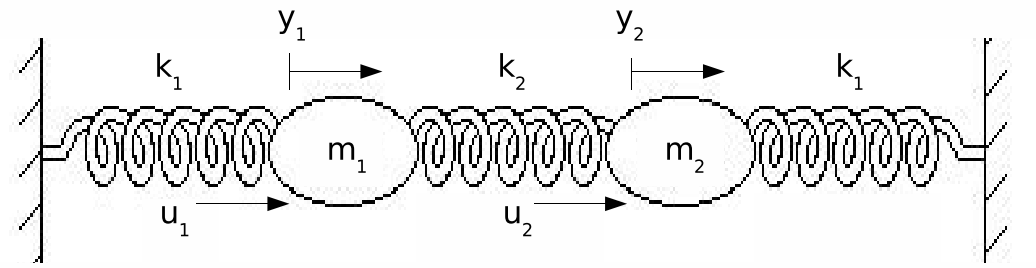


Figure 9.3: **Spring-mass system without friction.**

For the frictionless system in Fig. (**??**), $c = 0$. The input is given by the forces $u = [u_1, u_2]^T$, the output by the resulting displacement $y = [y_1, y_2]^T$.

The two-dimensional output $y_1$,$y_2$ represents the displacements of the two masses.

We calculated reduced realizations for a set of different $\beta$-values. The different reduced realizations correspond to points in the information plain spanned by $I(u_{past}, x)$ and $I(x, y_{fut})$. The coordinates can be calculated as follows:

$$I(u_{past}, x) = \log |H(\beta)H^T(\beta) + I|$$
$$I(x, y_{fut}) = I(u_{past}, x) - \log |H(\beta)H^T(\beta) - H(\beta)H^T(HH^T + I)^{-1}HH^T(\beta) + I|$$

The points for some realizations are represented as red crosses in **??**. The theoretical information curve as given by Equ. (**??**) is displayed as gray line in **??**. We see that all sample realizations lie on the optimal information curve. This is expected by construction and confirms the numerical implementation.

The Hankel matrices have finite dimension – for numerical purposes. The dimension scales with time $T$, the time window of past and future that are correlated, comparable to (**?**). In particular, a system with friction is correlated only over finite time, the information curve levels off, solid line. A system without friction has eigenvalues on the unit circle, the Hankel matrix does not decay with time. Hence, the information curve does not level off as a function of $T$ (Figure **??**).

## 9.6 Discussion

We will summarize the results of this chapter in a procedure that implements system identification and model reduction based on the information bottle-

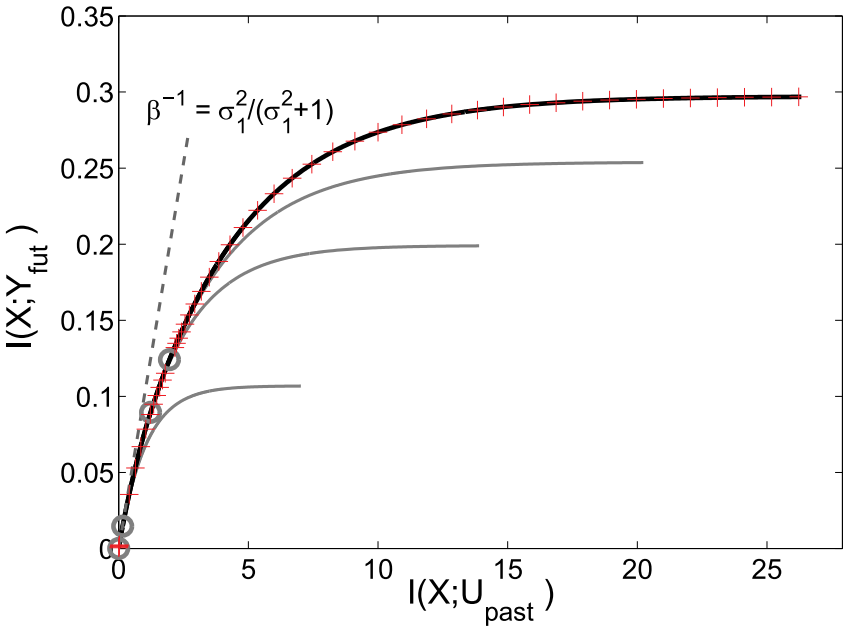


Figure 9.4: **Information curve for the spring-mass system.** The dynamics are given by the matrices in Equ. (**??**) and the following parameters: $m_1 = 5$; $m_2 = sqrt15$; $k_1 = 1$; $k_2 = 0.5$; $c = 0$. This particular system has Hankel singular values $\sigma_1 = 0.3358$; $\sigma_2 = 0.3109$; $\sigma_3 = 0.2374$; $\sigma_4 = 0.2103$. The reduced system (red crosses) all lie on the information curve as given by the Hankel singular values. This demonstrates the numerical feasability of the past-future information bottleneck.

neck method. We assume the observation of input and output data streams in time.

1. Perform CCA on the covariance matrix between input past and output future and calculate the state space.

2. Obtain system matrices by regressing state space on input past and output future.

3. Compute the Hankel matrix via observability and controllability matrices.

4. Calculate the Hankel singular values.

5. Obtain the reduced system by modifying the Hankel singular values according to a given constraint.

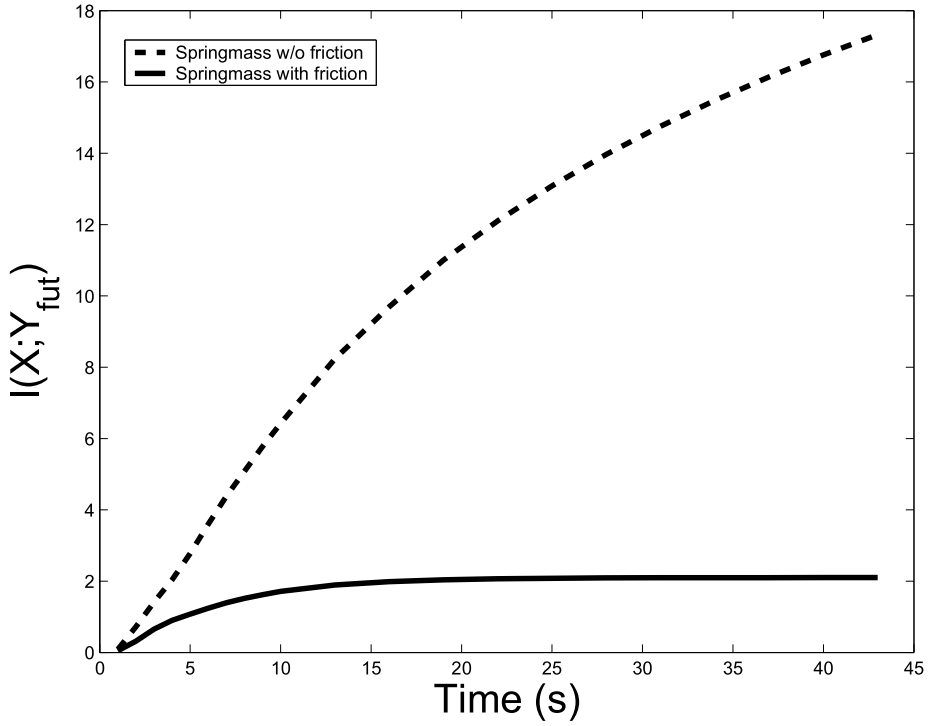


Figure 9.5: **Predictive information as a function of time.**

There are two possible perspectives. At the one hand, we may have a memory constraint, i.e., the information that we can keep about the past is limited. Via the information curve, we then obtain the maximal predictive information conditional on this limitation. On the other hand, we may impose a sufficiency requirement, stating that only a certain amount of predictive information is needed for a particular task. In this case, we can use the information curve to find the minimal memory capacity needed to achieve the required prediction accuracy. In fact, the information curve makes only sense if both dimensions, memory capacity and prediction accuracy, have some sort of soft or hard constraints or, equivalenty, cost/benefit functions. For example, storage costs could increase linearly with memory capacity, whereas the benefit of additional predictive information could level off.

From the perspective of an organism, the algorithmic procedure above is not needed. The animal might be interested in an internal model of some kind of external data stream. Its nervous system could simply perform PFIB directly on the past and future of the data stream, as in section (9.2), and adjust and weigh the individual state space dimensions according to $\sigma_i(\beta)$ or, equivalently, its associated canonical correlation coefficient.

From a technical point of view, results of this chapter are related to an immense quantity of literature. In fact, methods related to canonical correlation analysis of dynamical systems or time series can be found in the different fields of signal processing, machine learning, econometrics, neural networks, and dynamical system theory. Unfortunately, there is no consensus on notation and researchers are not always aware of the work in neighbouring field. Likewise, I cannot claim to have absorbed all or even the most part of the literature.

In a similar spirit to our results, the use of information between past and future for model selection has already been employed by calculating the information either with canonical correlation coefficients or by spectral densities (**????**) and can be traced back to **?**. **?** consider discrete-time stochastic processes, i.e., here the input is not entirely known. A maximum likelihood ansatz is used to derive system matrices for a finite data set. Other work shows that autoregressive moving average systems can be asymptotically efficient estimated by CCA and emphasizes that this approach provides accessible information on the appropriateness of the chosen model complexity (**?**). Furthermore, the canonical correlation coefficients estimated by CCA of past-future data were shown to be equal to the cosines of the principal angles between the linear subspaces spanned by input past and output future (**?**).

The difference of our work to all these results is that the information bottleneck allows a continous rather than a discrete tradeoff between two objectives and provides the computation of the optimal information curve. More fundamentally, the past-future information bottleneck can be seen as mapping a relationship between information theory and linear dynamical system theory.

Interestingly, a canonical correlation based approach can also be used for blind source seperation of mixed signals (**?**). This should come as no surprise with our results on the conceptual similarity between SFA and predictive coding (chapter 7) and another work, relating SFA and blind source separation (Sprekeler, personal communication).

### Summary and Outlook:

Neural systems need to encode temporally correlated data streams. As such data streams can be generally described as dynamical systems, the task can equivalently be phrased as a problem of identifying the underlying system. Furthermore, encoding must be sufficiently accurate for, e.g., appropriate behavioral output, but should avoid overaccurate representations. In the language of dynamical systems, this problem is called model reduction. Here,

we find the optimal information-theoretic tradeoff curve between accurate encoding of dynamical systems and permitted model reduction for linear dynamical systems. From this perspective, the state space can be seen as the information bottleneck between past and future of a data stream. System identification and model reduction by PFIB can be regarded as lossy source channel coding in the temporal domain and is shown to be similar to concurrent subspace-based system identification and balanced model truncation. The difference is that PFIB allows a continous tradeoff between model quality and complexity. We derive the relation to canonical correlation analysis of time series and calculate the information curve. We use the spring-mass system as an example to show that numerical simulations and theoretical predictions coincide. Fundamentally, this works shows that dynamical systems can be approached by information-theoretic methods.

The past-future information can and should be extended into several directions.

- Here, we compute the reduced system matrices via the PFIB-modified Hankel matrix. It would be interesting to obtain the reduced matrices directly from the orginal system. Plausibly, this can be achieved in a similar manner to balanced model truncation (**?**).

- Similar to the Gaussian information bottleneck and the past-future information bottleneck, channel capacity increases continously with weakening power constraints in the well-known case of water-filling for Gaussian channels (**??**). It is clear that the two approaches are closely related. Hence, working out the relation between water-filling in Gaussian channels and PFIB would connect the dynamical systems literature concerned with model reduction to many results from information theory.

- So far, we have applied the PFIB approach on deterministic systems. However, PFIB can also be adopted to stochastic systems where the system is (partially) driven by unknown noise. This would amount to additional estimation of the Kalman gain but should not interfere with general results.

- The assumption of this approach is stationarity of input and output signals. However, organism have to deal with changing, i.e., non-stationary environments. Hence, this formalism should be extended to adapt for changing statistics in an appropriate manner, as is done in neural systems (**?**).

- Speech processing is a suitable application for PFIB. Particularly, a comparison with biologically motivated algorithms relying on a version of predictive coding (**?**) seems appropriate.

- Local predictive coding is identical to linear SFA. How does (linear) SFA perform on non-Markovian processes in comparison to PFIB? Can the non-linear extension of SFA balance the missing information from previous time-steps?

- Neural ensembles encode information about the past and the future simultaneously in the hippocampus (**?**). In humans, imagining the future depends on much of the same neural machinery that is needed to remembering the past (**?**). A future challenge is the detailed understanding of these results in light of the past-information bottleneck.

- Finally, biological systems, especially neural networks, possess feedback as an essential component of their organization. An extension of PFIB to control design would be interesting and techniques from **?** could be used. For further discussion of this aspect, the reader is refered to Chapter 10.3.

# Chapter 10

# Outlook: Predictive coding in the grasshopper auditory system?

In the first part of this thesis, we have seen that the grasshopper *Chorthippus biguttulus* identifies particular temporal features of communication signals by a burst code in one ascending neuron. Also, integration of their spike trains allows time-scale invariant song identification. In the second part of this thesis, we have used an information-theoretic ansatz to find a tradeoff between accurate predictive coding and data compression in dynamical system. Here, we want to relate the two parts by asking: Is the grasshopper auditory system an information bottleneck for predictive information? First, we will extract the predictive components of grasshopper signals. Then we will discuss these results comprehensively in the context of the whole thesis and give an outlook for further research. Note that limitations and extensions of the different parts of this thesis are discussed at the end of chapters 5 and 9, respectively.

## 10.1 Predictive filters of grasshopper songs

Plausibly, the grasshopper relies on its auditory system for a variety of tasks. However, it is clear that at least communication signals have behavioral significance. Our past-future bottleneck approach (PFIB) suggests to ask the following question: Does the grasshopper auditory system extract those parts of communication signals that contain predictive information, i.e., information on the future components of the same signal? We will apply PFIB on grasshopper communication signals and discuss the results.

**Description of the algorithm.** All 8 communication signals were rectified and smoothed (2 ms) to give amplitude-modulation signals. Only the steady-state part of songs was chosen, as in chapter 2. Also, all songs were normalized with respect to mean and standard deviation. A sliding window with 200 sampling points was driven over the signals, extracting a vector with 200 entries. Each vector was defined as $X_{past}$; the associated vector with the subsequent 200 entries was defined as $X_{fut}$. Theorem **??** was applied on $X_{past}$ and $X_{fut}$, i.e., we calculated the left eigenvectors and eigenvalues of $\Sigma_{X_{past}|X_{fut}}\Sigma^{-1}_{X_{past}}$. The eigenvectors, i.e., the row vectors of matrix $A$ in **??** were interpreted as filters extracting those signal components that contain information about the future signal. For different filter time scales different sampling rates were chosen: 20 kHz, 6.6 kHz, 3.3 kHz, 1 kHz, 0.5 kHz and 0.2 kHz corresponding to filters with the following durations: 10 ms, 30 ms, 60 ms, 100 ms, 200 ms and 500 ms. The first 6 filters are displayed for the 10 ms, the 100 ms and the 200 ms time scale in Fig. **??**. Individual filters were applied on incoming signals to obtain associated state space trajectories. The first 6 trajectories are displayed for filters with time scale 100 ms (Fig. **??**).

**The filterbank.** First, we focus on the filters with short time scale, i.e., 10 ms. Most predictive information is on very short time scales. The first filter with time scale 10 ms already contains most information on the future, relying basically on the direction of amplitude change within the last 0.1 ms (Fig. **??**a)[1]. However, from the inspection of the subsequent filters, we can assert that amplitude modulations of 2 ms are also relevant (Fig. **??**a). In contrast to a Fourier decomposition, the filters have no sinus-shape. The second filter emphasizes periodically appearing mini amplitude plateaus (1.6 ms) interrupted with mini gaps (0.4 ms). The third filter extracts periodic mini onsets preceded by equally short excursion below average amplitude. As a second example, we investigate the filters with time scales of 100 ms (Fig. **??**b). The first two filters extract fluctuations on the 2 ms time scale, similarly to the set of filters above. However, subsequently, predictive components on longer time scales appear. The third, forth and sixth filter show unregular amplitude modulation at a timescale of around 25 ms. Third, we also show the 6 first filters for the 200 ms time scale. Also here, the first two filters extract very short time scale flucutations. The third filter has a rather slow modulation. The forth, fifth and sixth filter extract signals with

[1]The preceding fluctuations probably occur due to forced orthogonalization with the other filters.

a periodicity of around 50 ms.

**State space trajectories.** Filters extract predictive information from temporal patterns and project this information into state space trajectories. The projection of one specific communication signal by individual filters is displayed in Fig. (**??**). The first filter alone is sufficient to reprocude the complete song dynamics. Basically, this filter reproduces the amplitude 0.05 ms ago. Hence, this result should not be very surprising. The subsequent filters at 10 ms seem to be much less predictive, as shown in Fig. (**??**B). However, this is due to the fact that everything is already predicted by the first filter.

The first filter at 100 ms with 500 Hz components, i.e. amplitude modulations at 2 ms, can also correctly predict the time course of the communication signal (Fig. **??**C). The state space trajectory of the third filter is interesting as its dynamics are already much slower (Fig. **??**D). Furthermore, the amplitude modulation can be seen as a phase-shifted abstraction of the detailed syllable-pause alternation, i.e., phase information is not necessarily preserved. The sixth filter of Fig. (**??**b) is truely predictive, as it produces a trajectory preceding the communication signal structure with several milliseconds (Fig.**??**E). The very slow third filter for the 200 ms filter set (Fig. **??**c) can be interpreted as a further abstraction, only signalling syllable pause alternations with no respect of fine-detailed temporal structures (Fig. **??**F). On the other hand, also this time scale has filters with faster components associated with more detailed state space trajectories (Fig. **??**G).

**Discussion.** What can we learn from these observations? There seem to be different time scales within the communication signals that carry predictive information. Of course, the immediate past is highly predictive about the signal (Fig. **??**A). Beside this not surprising result, periodic components at 2 ms (Fig. **??**B+C), at 25 ms (Fig. **??**D+E), at 50 ms (Fig. **??**G) and longer time scales (Fig. **??**F) carry information about the future signal.

It would be interesting to compare individual filters with cell properties in the grasshopper auditory system. Here, we just want to note some similarities. Short time fluctuations of around 2 ms are quite well represented at the level of receptor neurons, where the maximal firing rate is ca. 500 Hz. Also, stimulus reconstruction based on spike trains of receptor neurons is quite accurate (**?**). Ascending neurons, on the other hand, seem to be less well suited for accurate stimulus reconstruction as i) maximal spike rate is cut in half from receptor to ascending neurons and ii) interspike-interval variability increases from receptor to ascending neurons (**?**). In detail, the AN6 neuron is the only ascending neuron responding tonically to amplitude modulations but encoding song patterns much less precisely than receptor neurons (**?**).

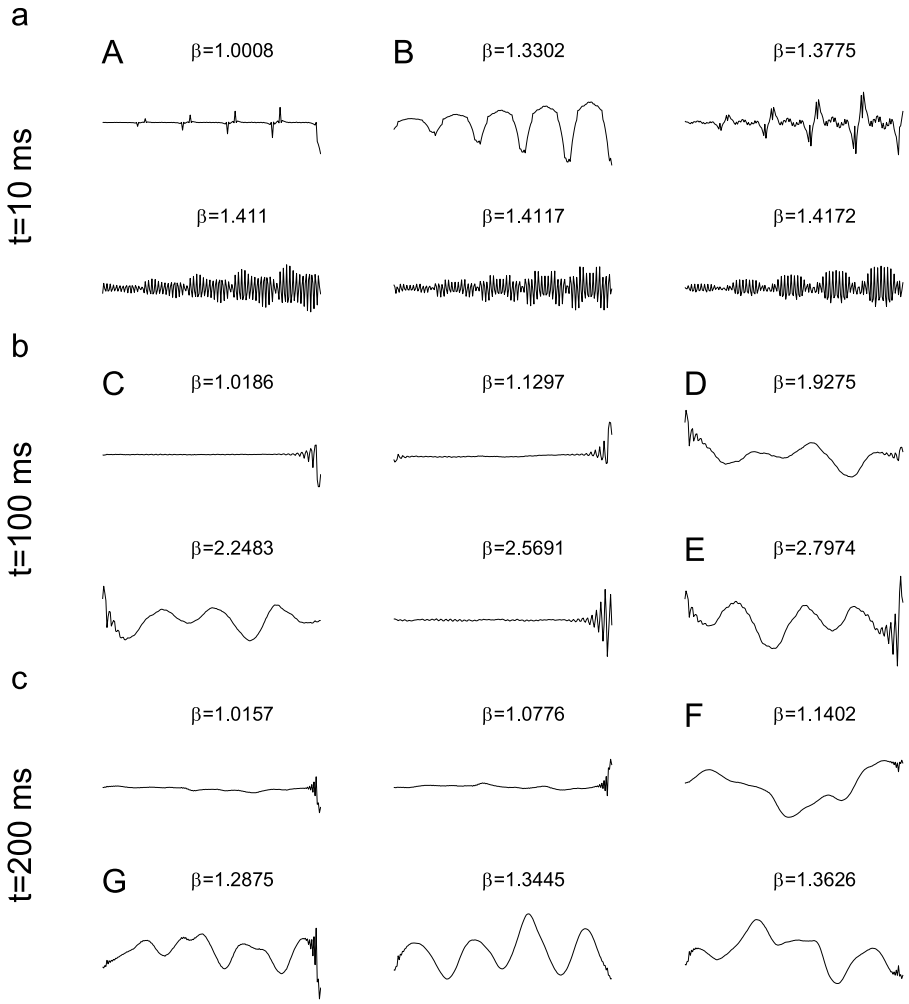


Figure 10.1: **Predictive filters of grasshopper communciation signals for different time scales.** For each time scale, the 6 most predictive filters are displayed. According to time scale, filters in a) b) and c) are 10 ms, 100 ms and 200 ms long. The capital letters indicate the filters used for Fig. (**??**).

However, modulations on longer time-scales can be represented. As we have seen in chapters 2–5, the AN12 can encode the pause duration. Also, the AN6 fires tonically in response to syllables, thereby encoding syllable duration (**?**). These coding properties are reminiscent of filters with longer time-scales such in Fig. (**??**D-F).

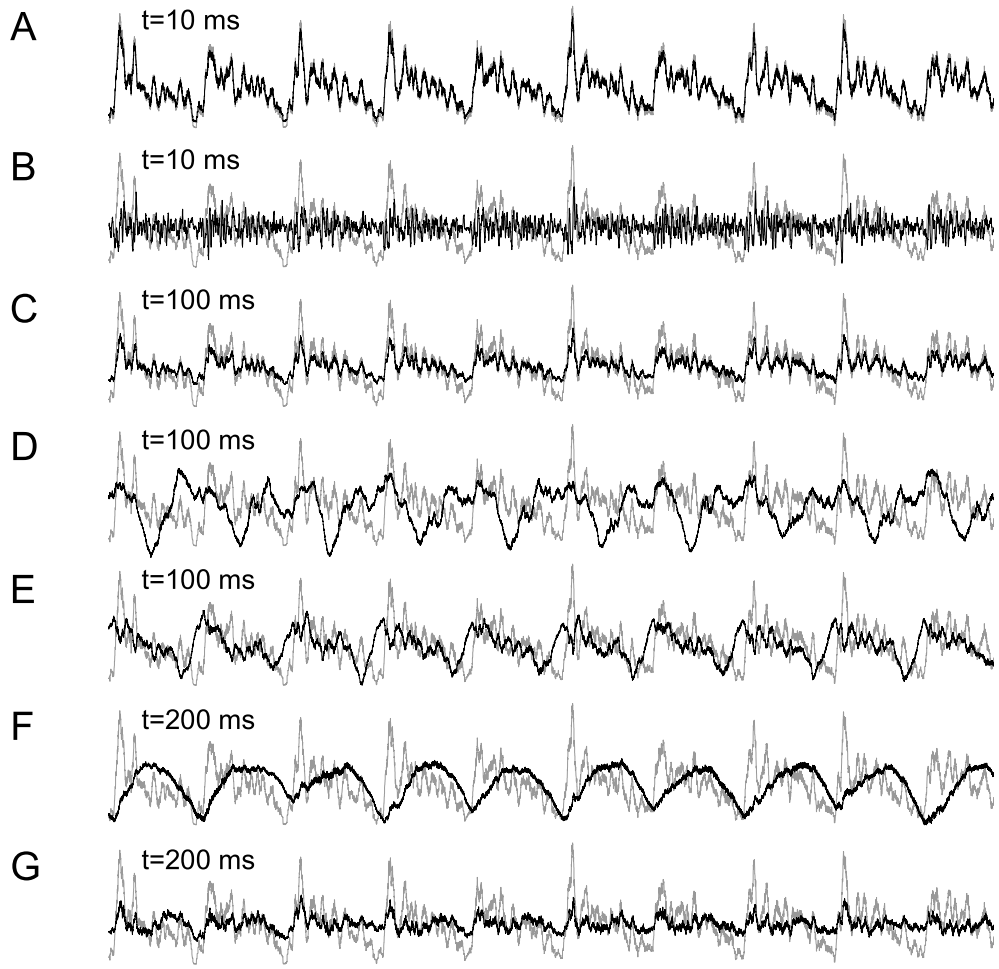


Figure 10.2: **State-space trajectories for communication signals.** Capital letters correspond to indicated filters in Figure **??**. The input signal is in gray, state space trajectories in black. See text for detailed discussion.

## 10.2 The auditory system as an information bottleneck

In one regard, the past-future information bottleneck approach is clearly limited. The predictive filters were extracted while ignoring the architecture of

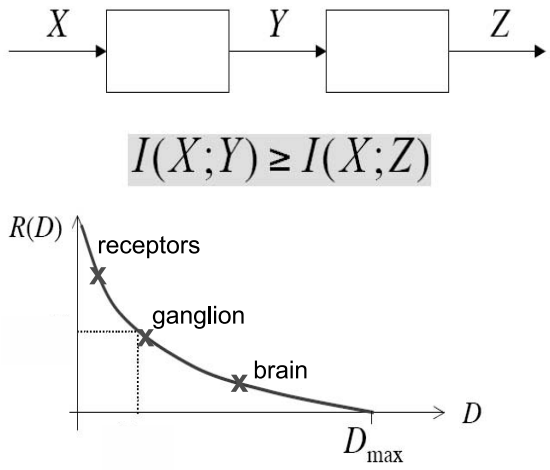


Figure 10.3: **Data processing inequality along the auditory pathway.** According to the data processing inequality, the information $Z$ contains about $X$ cannot be larger than $I(Y, X)$ if $Z$ is a function of $Y$ only. Hence, in hierarchically organized nervous systems higher processing stages contain less information about the input than lower ones: the feasability of reconstruction worsens. The rate-distortion curve shows the minimal distortion that can be achieved when information rate is reduced.

the sensory system. In fact, the latter imposes interesting boundary conditions.

A main attribute of the sensory system is its hierarchical organization (Fig. **??**). By the data processing inequality, interneurons cannot carry more information about the signal than the joint activity of receptor neurons do, and similarly ascending neurons and subsequent read-out cells carry less and less information (Fig. **??**). What kind of information processing strategy is appropriate for such a feed-forward network?

First, one may argue that each processing stage should try to keep effectively all information of the preceding stage about the signal, approximating a data processing equality. However, this is not necessarily reasonable or possible. For example, a converging architecture might limit the upstream information rate. Second however, the reduction in information rate may still follow an optimality principle, such as the rate-distortion curve (see chapter 6). Upstream neurons would keep that information, that allows best possible reconstruction (smallest distortion) or decoding given the limited information rate (Fig. **??**). One very common distortion measure is the mean square distance between reconstruction and signal. In fact, the grasshopper receptor neurons can accurately reproduce the stimulus time course with low mean square error and high information rate: 500 bits/s (**?**). We can also guess that at least single ascending neurons, having a lower firing rate and

probably lower information rate $< 100$ bits/s, can reconstruct stimulus time course with lower quality. In fact, the AN12 may signal the timing of syllable onset and pause duration, hence, forwarding the slow time-scale modulations of the signal. However, the mean square error may actually prove to be a meaningless distortion error. The AN12 neuron may reproduce some aspects of the temporal patterns but it may also only poorly encode absolute signal amplitude, potentially producing a large mean square distance.

There is another but related characteristic of this hierarchical auditory sytem. Our observations indicate that higher processing stages integrate information over subsequently longer time scales. Receptor neurons fire in response to amplitude modulations of 2 ms, ascending neurons prefer time scales of 10-70 ms. A plausible read-out neuron integrates over 0.5-1 s. Hence, higher level neurons are more sensitive to modulations on successively longer time scales. This observation cannot be deduced from the predictive filters in the last section. There, predictive filters appear simultaneously.

Indeed, integration of long time-scales might be a common property of auditory system. As discussed in Chapter 3, also neurons in auditory cortex of mammals integrate over a variety of timescales (10 ms, 100 ms, 1 s) (**?**). The example of the AN12 neuron (Chapter 2), results of our modeling studies (Chapters 3 + 5) and this chapter's considerations of the predictive coding hypothesis indicate that invariant auditory object recognition of temporal patterns may be achieved by temporal integration over successively longer time-scales.

What then is the function of the auditory system of the grasshopper? Filter properties of different levels are reminiscent of theoretical predictive filter. Hence, low level information processing is not in contradiction to the hypothesis that the sensory system extracts predictive information. The firing rate of receptor neurons can be used for accurate stimulus reconstruction but it is not clear that higher stages can perform similar tasks. That is, a rate-distortion theory with a simple distortion measure can probably not be used to model the auditory system. In contrast, higher processing stages seem to extract invariant temporal features on longer time scales. Hence, we can postulate the following hypothesis:

> *The auditory system extracts predictive information such that neural states of higher hierarchical levels contain information about longer time scales than the neural states of lower levels.*

Note that this hypothesis already implies that total information rate decreases along the auditory pathway as short time predictions are more infor-

mative than long time predictions (Fig. **??**+**??**+**??**). On the other hand, it is information on longer time scales ($\sim$ reaction time) that is behaviorally relevant – and hence sufficient – for an animal.

From this hypothesis, we can derive questions for experimental and computational studies:

- Are auditory neurons tuned to extract predictive statistics of the signal? How do response properties change if the content of predictive information in signals is altered? How does the code change with the level of hierarchy and integration time?
- Can the auditory system of insects, songbirds or mammals be modeled by a hierarchy of predictive filters? Can existing models (**?**) be extended into hierarchical (non-linear) models similar to (**??**) but with a PFIB objective function?

Such models should be seen in context of established frameworks of object recognition in the visual system. For example, invariant visual object recognition algorithms suggest that high-level invariances are detected by successive spatial integration over low-level features (**??**). The close relationship of predictive coding and slow feature analysis, an established framework for invariance detection (**?**), as shown in chapter 7, indicates that the past-future information bottleneck can be the basis of a hierarchical network detecting invariances in the temporal domain.

## 10.3 Relevant signals and self-referential systems

We can proceed one step further. From behavioral experiments we can deduce that the presence of a mating song corresponds to one top-level invariance: Grasshoppers respond stereotypically to communication signals of sufficient quality. Of course, such a communication signal itself is predictive on a higher time scale, i.e., indicating reproduction in the near future. From this point of view it becomes even more clear that sensory systems extract not all predictive structures but focus on those that have predictive and relevant value for their own future.

However, the notion of behavioral relevancy imposes conceptual difficulties. How should one define *relevant* features? The grasshopper communication signals are relevant because other grasshoppers respond to them. And other grasshoppers respond to them because the communication signals are behaviorally relevant. Hence, we encouter a recursive definition of *relevancy*.

From the empirical point of view this poses no further problems as one is – as an empirical scientist – first of all interested in an accurate description. In contrast, for a theoretician it is less clear how to formalize the underlying logic of such autopoietic processes.

The foundations of a theory of autopoiesis in biological systems have been laid out vividly by **?**. In his terms, an autopoietic system is organized as a network of processes of production of components that – through their interactions – continously regenerate and realize the network that produces them. From this perspective, organisms are autopoietic (literally *self-produced*) systems with constantly changing structure (e.g., protein concentrations) by maintaining its organization. The nervous system itself can be regarded as such an autopoietic system. The nervous system is a closed network of interacting neurons such that the change of activity in some neurons lead to change in activity in other neurons, either directly through synaptic action or indirectly through genetic coupling or physical mediators in the environment. From this perspective, effector neurons change the environment such that the activity level of sensory neurons is modified. The fundamental invariance of the nervous system is then the maintenance of the relations that define its participation in the higher-order autopoietic system, the organism. Also sensory perception can be interpreted as the construction of invariances through sensory-motor coupling. For example, in locomotion a rhythmic pattern generator generates motor output. Proprioceptors sense the output and give feedback to the pattern generator such that the walking rhythm of the animal is invariant to minor environmental changes. How does such a perspective interfer with the approach of this thesis? Varela claims that – switching from an information-theoretic or engineering to an autonomy perspective – *every bit of information is relative to maintenance of a system's identity, and can only be described in reference to it, for there is no designer*. Varela framed the information-theoric input-output and the autonomous point of view as antithetic, or as complementary. However, the notion of *relevant* or *sufficient* information, as used in (**?**) and within this thesis, instead suggests that the two perspectives can be combined. The relevancy and sufficiency of information is then evaluated with respect to the maintenance of the organism.

A practical way to deal with autopoietic systems can be found in evolutionary theories formulated with a variety of agents and reinforcement learning algorithms on different interacting time scales. Hence it is obvious, that these dynamics are fundamentally shaped by feedback processes on all levels. The past-future information bottleneck is reasonably motivated but – as a first approach – restricts itself on processes without feedback. Fortunately, dynamical system theory can naturally be extended to control theory

that includes feedback loops into the framework, e.g. (**?**). The focus, then, should also shift from input-output relations to construction of invariances of internal states. A suitable basis for this is specified by behavioral model identification where all dynamics are put into an autonomous state-space model – instead of an input-output model (**?**).

## 10.4 Closing words

What are principles of sensory processing of temporal signals? Both efficiency and relevancy arguments suggest predictive coding as a guiding principle. This claim is substantiated by our result that local predictive coding and an established computation model of sensory processing – slow feature analysis – are equivalent under certain conditions. We demonstrate that the information-theoretic perspective can be combined with the engineering dynamical system formalism. The problem of finding an internal state that maximizes predictive coding while keeping overall information rate low is mapped onto a joint system identification / model reduction problem, enabling sufficient coding. Changing to the neural perspective, this thesis also demonstrates that the auditory system can transform temporal signal features such as pause durations into a graded intraburst spike count code. Rather than forwarding this information, subsequent brain neurons could directly read-out communication signals in a time-scale invariant manner by integration. This is concordant to the view that the auditory system extracts predictive filters of relevant stimulus statistics at successively longer time-scales.

# Appendix A

# Analysis of the bursting interneuron: Methods

**Stimulus design.** Two different stimulus schemes were used: Natural mating songs (recordings from Sandra Wohlgemuth) and artificial block stimuli (recordings from Andreas Stumpner). Under the natural song scheme, acoustic search stimuli were presented to identify auditory neurons. Then a brief intensity response scheme was run to determine the neuron's response characteristics (100 ms pulses filled with white noise, bandwidth 0.5-30 kHz, 30-70 dB in 10-dB steps, each intensity repeated four times). The acoustic stimuli used in the natural song stimulus set were eight different songs (Fig **??**a) recorded from individual males all of which were able to evoke a positive female response. Each song was repeated eight times while intracellularly recording the response of the neuron.

The acoustic stimuli used in the artificial song scheme consisted of six syllables of rectangularly modulated white noise ($2.5 - 40$ kHz, Fig **??**b). Longer versions of the same models were used in previous behavioural experiments (e.g. **?**). The six syllables had constant durations (40, 85 or 110 ms, the pauses in between were 3.2, 8, 8, 16.3, 24.0, 33.0 and 42.5 ms long. Each stimulus set was repeated 8 times.

**Animals, electrophysiology and acoustic stimulation.** Experiments under the natural song stimulus paradigm were performed on 6 animals, 3 *Chorthippus biguttulus*, and 3 adult locusts (*Locusta migratoria*). In both auditory systems the same kind of auditory neurons, specifically ANs can be anatomically identified and electrophysiologically characterized (**??**); furthermore, interspecific spike train distance between AN12 cells is similar to intraspecific spike train distance (**?**). During the experiments with natural songs the preparations were kept at a constant temperature of $30 \pm 2^{\circ}$

C. Details of the experimental procedure are given in Wohlgemuth et al., (2007).

Experiments under the artificial block stimuli paradigm were performed on 9 adults of *Chorthippus biguttulus*. The experiments were conducted at a constant temperature of $25 \pm 2°$ C. The electrophysiological methods and stimulution apparatus were similar to the natural song stimulus set and are described in detail in (**?**). The spike count was much lower than in the natural stimulus paradigm, probably due to the lower recording temperature (25° C vs 30° C).

**Data analysis.** From the digitized recording signal, spike times were determined by means of a voltage threshold criterion. The first part of each song was dismissed to include only the steady-state part of the songs into the analysis. For spike train analysis, any cluster of spikes was defined as a burst. Formally, we included single spikes and treated them as bursts. A spike belongs to a burst only if the spike follows the preceding spike after not more than $3 + n$ ms, $n$ being the spike rank within the burst. This measure takes into account that interspike intervals tend to increase with ongoing burst duration. Altogether, we dismissed information in the temporal fine structure within bursts and instead concentrated on spike count information only.

In natural songs, each syllable onset is preceded by a period of relative quietness. In fact, the amplitude-modulation signal during pauses is not vanishing as under the artificial song scheme (Fig. **??**). This difference is not negligible (**?**) but does not effect our results in principle. Hence, for simplicity, the period of relative quietness in natural songs will also be called "pause". The pause in natural songs is defined as the time between passing a certain amplitude threshold from above and passing this threshold again from below (Fig. **??**c). For each cell, the songs were normalized with respect to mean and standard deviation. The threshold was varied such that the correlation between pause and spike count within the subsequent burst was maximized. The correlation is robust with respect to different amplitude levels (Fig. **??**c). The minimal amplitude is taken as the minimal value in the preceding period of quietness (Fig **??**c). Relative onset amplitude is defined as the maximal onset amplitude minus the minimal amplitude in the preceding quietness period. The total period duration is defined as the time difference between the first spikes of two subsequent bursts (Fig **??**d). The slope is defined as the relative onset amplitude divided by the time interval between maximal and minimal amplitude values (Fig **??**c). To obtain the correlation between spike count and signal features, the R-square value (explained variance, Pearson correlation) was calculated. It was checked for

multicollinearity by computing the semi-partial r-value for each independent variable.

**Song classification**. The classification of AN12 responses is based on intraburst spike count only, i.e., no temporal information is used. For the classification, we used only those bursts which had, on average over 8 trials, more than 1 spike at any specific time, thus including only reliable events. The order of bursts within a song was indexed such that the $k$-th burst has burst index $k$. The average number of intraburst spike count is assigned to the corresponding burst index. The spike count difference between each intraburst spike count to all other intraburst spike counts at the same burst index is measured. This intraburst spike count is then assigned to that song whose intraburst spike counts have in average the smallest difference. This classification scheme was first applied to individual bursts. Second, cumulutive classification based on $n$-bursts was done by adding the spike count differences of $n$ subsequent bursts before classification.

**Estimation of Mutual Information.** We used the adaptive direct method developed by **?** to calculate the mutual information (MI) between pause duration and spike count within bursts. Here, one starts with the best resolution available and constructs the matrix of joint probabilities between spike counts and pause durations. The naive MI and the bias corrected MI are computed using the method from Panzeri and Treves (**?**) for bias correction. Step by step, one reduces the matrix by merging columns which represent the finely binned pause durations. The matrix was reduced by merging the column with the smallest marginal probability with that neighbouring column that has the smaller marginal probability. The result is a set of decreasing MI and corresponding bias values. The true MI is estimated as the largest difference between those two values.

**Mutual information between cumulated spike count and song identity.** The probability of correct classification was used to calculate the mutual information between cumulated burst events and song identity. We assumed that wrong classifications were equally distributed across the other 7 songs, thus giving a strict lower bound on the mutual information. By this, only the relevant information of correct classification was taken into account.

# Appendix B

# Derivation of the generalized eigenvalue equation for SFA

Let $W_j$ denote the row vector that is formed by the $j$-th row of the weight matrix $A$. The output signal $Y_j$ is then given by $Y_j = W_j X$. Accordingly, the slowness objective (**??**) is given by

$$\begin{aligned} \Delta(Y_j) = & \quad \langle \dot{Y}_j^2 \rangle_t \\ = & \quad \langle (W_j \dot{X})(W_j \dot{X}) \rangle_t \\ = & \quad W_j \langle \dot{X} \dot{X}^T \rangle_t W_j^T = W_j \Sigma_{\dot{X}} W_j^T \end{aligned} \tag{B.1}$$

A similar calculation yields that the variance of the output signal $Y_j$ is given by

$$var(Y_j) \equiv \langle Y_j^2 \rangle_t = W_j \langle X X^T \rangle_t W_j^T = W_j \Sigma_X W_j^T \overset{(??)}{=} 1 \,. \tag{B.2}$$

The task is to minimize (**??**) under the constraint (**??**) and the decorrelation constraint, which we will neglect for now as it will turn out to be fulfilled automatically. The method of Lagrange multipliers states the necessary condition that

$$\Psi = \Delta(Y_j) - \lambda \langle Y_j^2 \rangle_t$$

is stationary for some value of the Lagrange multiplier $\lambda$, i.e., that the gradient of $\Psi$ with respect to the weight vector $W_j$ vanishes. Using (**??**) and (**??**), this gradient can be calculated analytically, yielding the following necessary condition for the weight vector $W_j$

$$W_j \Sigma_{\dot{X}} - \lambda W_j \Sigma_X = 0 \,. \tag{B.3}$$

Note that condition (**??**) has the structure of a generalized eigenvalue problem, where the Lagrange multiplier $\lambda$ plays the role of the eigenvalue. Multiplication with $W_j^T$ from the right and using the unit variance constraint (**??**)

yields that the $\Delta$-value of a solution of (**??**) is given by its eigenvalue $\lambda$:

$$\underbrace{W_j \Sigma_{\dot{X}} W_j^T}_{\stackrel{(??)}{=}\Delta(Y_j)} - \lambda \underbrace{W_j \Sigma_X W_j^T}_{\stackrel{(??)}{=}\langle Y^2 \rangle_t = 1} = 0 \quad \Rightarrow \quad \Delta(Y_j) = \lambda \,.$$

From this it is immediately clear that the slowest possible output signal is provided by the linear function associated with the eigenvector $W_1$ with the smallest eigenvalue $\lambda_1$. It can be shown that eigenvectors $W_i, W_j$ with different eigenvalues $\lambda_i, \lambda_j$ are orthogonal in the sense that $\langle Y_i Y_j \rangle_t = W_i \Sigma_X W_j = 0$, so they yield decorrelated output signals. For eigenvectors with identical eigenvalues, any linear combination of them is still an eigenvector. Hence, it is always possible to choose a basis of the subspace that still consists of eigenvectors and yields decorrelated output signals (e.g. by Gram-Schmidt orthogonalization).

Combining these properties of the eigenvectors, it is clear that the optimization problem of linear SFA can be solved by choosing the functions associated with the $J$ eigenvectors $W_j$ with the smallest eigenvalues, ordered by their eigenvalue. Reinserting the eigenvectors $W_j$ into the matrix $A$ and the eigenvalues in a diagonal matrix $\Lambda$, the eigenvalue problem (**??**) takes the form of equation (**??**)

$$A \Sigma_{\dot{X}} = \Lambda A \Sigma_X \,.$$

# Appendix C

# Derivation of the Optimal Weight Matrix for Local Predictive Coding

We first rewrite the mutual information quantities in the objective function for local predictive coding in terms of differential entropies:

$$\begin{aligned} \mathcal{L}_{LPC} = & \quad I(Y_t, X_y) - I(Y_t, X_{t+1}) \\ = & \quad h(Y_t) - h(Y_t|X_t) - \beta h(Y_t) + \beta h(Y_t|X_{t+1})\,. \end{aligned} \tag{C.1}$$

Here, the differential entropy of a stochastic variable $Z$ is given by $h(Z) = -\int_Z f(z)\log f(z)dz$ with $f(z)$ denoting the probability density of $Z$. In particular, for Gaussian variables, the differential entropy becomes

$$h(Z) = \frac{1}{2}\log\left(2\pi e\right)^d|\Sigma_Z|\,,$$

where $|\Sigma_Z|$ denotes the determinant of $\Sigma_Z$ and $\Sigma_Z := \langle ZZ^T\rangle_t$ is the covariance matrix of $Z$ (**?**). Hence, we have to find the covariance matrices of the quantities in (**??**). As $Y_t = AX_t + \xi$, we have $\Sigma_{Y_t} = A\Sigma_{X_t}A^T + \Sigma_\xi$ and $\Sigma_{Y_t|X_t} = \Sigma_\xi$. The last covariance matrix is obtained as follows:

$$\begin{aligned} \Sigma_{Y_t|X_{t+1}} &= \Sigma_{Y_t} - \Sigma_{Y_t;X_{t+1}}\Sigma^{-1}_{X_{t+1}}\Sigma_{X_{t+1};Y_t} \\ &= A\Sigma_{X_t}A^T + \Sigma_\xi - A\Sigma_{X_t;X_{t+1}}\Sigma^{-1}_{X_{t+1}}\Sigma_{X_{t+1};X_t}A^T \\ &= A\Sigma_{X_t|X_{t+1}}A^T + \Sigma_\xi\,, \end{aligned}$$

where we used Schur's formula, i.e. $\Sigma_{X|Y} = \Sigma_X - \Sigma_{X;Y}\Sigma_X^{-1}\Sigma_{Y;X}$, in the first and last step (**?**). Neglecting irrelevant constants and using that the noise is isotropic, the objective function (**??**) becomes

$$\mathcal{L} = (1-\beta)\log|A\Sigma_{X_t}A^T + I| + \beta\log|A\Sigma_{X_t|X_{t+1}}A^T + I|\,. \tag{C.2}$$

The derivative of the objective function with respect to the weight matrix is given by

$$\frac{d\mathcal{L}}{dA} = (1-\beta)(A\Sigma_{X_t}A^T + I)^{-1}2A\Sigma_{X_t} + \beta(A\Sigma_{X_t|X_{t+1}}A^T + I)^{-1}2A\Sigma_{X_t|X_{t+1}}\,.$$

Equating this to zero and rearranging, we obtain a necessary condition for the weight matrix A:

$$\frac{\beta-1}{\beta}\underbrace{(A\Sigma_{X_t|X_{t+1}}A^T + I)(A\Sigma_{X_t}A^T + I)^{-1}}_{=:M}A = A\Sigma_{X_t|X_{t+1}}\Sigma_{X_t}^{-1}\,. \tag{C.3}$$

We will prove that this equation can be solved by filling the rows of $A$ with adequately scaled versions of the solutions $W_j$ of the following generalized (left) eigenvalue problem:

$$W_j\Sigma_{X_t|X_{t+1}} = \lambda_j W_j\Sigma_{X_t}\,. \tag{C.4}$$

We will first make some considerations on the solutions of the eigenvalue equation (**??**) and then insert them into equation (**??**) to show that this yields $M$ diagonal. It then becomes clear that there are scaling factors for the eigenvectors such that equation (**??**) is solved.

(1) $W_j$ **is a left eigenvector of** $\Sigma_{X_t|X_{t+1}}\Sigma_{X_t}^{-1}$:

$$\begin{aligned} W_j\Sigma_{X_t|X_{t+1}} &= \lambda W_j\Sigma_{X_t} \\ \Leftrightarrow W_j\Sigma_{X_t|X_{t+1}}\Sigma_{X_t}^{-1} &= \lambda W_j\,. \end{aligned} \tag{C.5}$$

(2) $M$ **is diagonal:** The crucial observation for this statement is, that the eigenvectors $W_j$ need not to be orthogonal, because $\Sigma_{X_t|X_{t+1}}\Sigma_{X_t}^{-1}$ is not necessarily symmetric. The structure of the generalized eigenvalue equation is such that solutions of equation (**??**) with different eigenvalues $\lambda$ are orthogonal with respect to the positive definite bilinear form induced by $\Sigma_{X_t}$:

$$(W_i, W_j) = W_i\Sigma_{X_t}W_j^T = r_i\delta_{ij} \quad \text{with} \quad r_i > 0\,.$$

In the case where there are several eigenvectors with the same eigenvalue, it is always possible to choose eigenvectors $W_i$ that are orthogonal in the sense above. Assume that the rows of $A$ are filled with the eigenvectors $W_j$, scaled by a factor $\alpha_j$. With this choice, $A\Sigma_{X_t}A^T + I$ is diagonal with diagonal elements $r_j\alpha_j^2 + 1$. Right multiplication of (**??**) with $W_j^T$ yields that $A\Sigma_{X_t|X_{t+1}}A^T + I$ is also diagonal with diagonal elements $r_j\lambda_j\alpha_j^2 + 1$. Thus $M$ is diagonal with diagonal elements $M_{jj} = \frac{r_j\alpha_j^2\lambda_j+1}{r_j\alpha_j^2+1}$.

(3) Using the above results, (**??**) becomes

$$\left[\frac{\beta-1}{\beta}\frac{\lambda_j\alpha_j^2 r_j+1}{\alpha_j^2 r_j+1}-\lambda_j\right]\alpha_j W_j=0\,. \tag{C.6}$$

This equation can only solved if either $\alpha_j = 0$ or

$$\frac{\beta-1}{\beta}\frac{\lambda_j\alpha_j^2 r_j+1}{\alpha_j^2 r_j+1}=\lambda_j\,.$$

Rearranging for $\alpha_j^2$ yields the normalization stated in proposition 1:

$$\alpha_j^2=\frac{\beta(1-\lambda_j)-1}{\lambda_j r_j}\,.$$

Of course this equation can only be solved if the right hand side is positive. Because $r_j$ and $\lambda_j$ are positive, this reduces to a relation between the $\beta$-value and the eigenvalues:

$$\beta\geq\frac{1}{1-\lambda_j}\,.$$

For the eigenvalues that do not fulfill this condition for a given $\beta$, equation (**??**) can only be solved by $\alpha_j = 0$. This shows that the critical $\beta$-values as stated in proposition 1 are those, where a new eigenvector becomes available. Moreover, we have now demonstrated that $A(\beta)$ as stated in proposition 1 is a solution of equation (**??**). Note that in line with the fact that the objective function of optimization problem 1 is invariant with respect to orthogonal transformations of the output signals, any matrix $\tilde{A} = UA$ with $U^{-1} = U^T$ is also a solution of (**??**). We refer the reader to (**?**) for the proof that $A(\beta)$ is not only a stationary point of (**??**) but also minimizes the objective function (**??**).

# Bibliography

M. Abeles. *Corticonics: Neural Circuits of the Cerebral Cortex*. Cambridge University Press, Cambridge, United Kingdom, 1990.

E. D. Adrian. *The Basis of Sensation: The Action of the Sense Organs*. W. W. Norton, New York, 1928.

H. Akaike. *Canonical correlation analysis of time series and the use of an information criterion*, pages 27–96. (Mehra, R. and Lainiotis, D., Eds.) Academic, 1976.

T. B. Alder and G. J. Rose. Long-term temporal integration in the anuran auditory system. *Nat. Neurosci.*, 1:519–522, 1998.

H.J. Alitto, T.G. Weyand, and W.M. Usrey. Distinct properties of stimulus-evoked bursts in the lateral geniculate nucleus. *J. Neurosci.*, 25:514–523, 2005.

C. Allen and C. F. Stevens. An evaluation of causes for unreliability of synaptic transmission. *Proc. Natl. Acad. Sci.USA*, 91:10380–10383, 1994.

J. S. Anderson, I. Lampl, D. C. Gillespie, and D. Ferster. The contribution of noise to contrast invariance of orientation tuning in cat visual cortex. *Science*, 290:1968–1972, 2000.

M. Aoki. Control of large-scale dynamic systems by aggregation. *IEEE Tran. on Autmatic Control*, 13:1–2, 1968.

K.J. Åström and T. Bohlin. Numerical identification of linear dynamic systems for normal operating records. *Proc. 2nd IFAC Symp. Theory of Self-Adaptive Systems*, pages 96–111, 1965.

Jospeh J. Atick. Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, 3:213–251, 1992.

F. Attneave. Some informational aspects of visual perception. *Psychol. Rev.*, 61:183–193, 1954.

R. Azouz, M. S. Jensen, and Y. Yaari. Ionic basis of spike after-depolarization and burst generation in adult rat hippocampal ca1 pyramidal cells. *J. Physiol.*, 492:211–223, 1996.

R. Balakrishnan, D. von Helversen, and O. von Helversen. Song pattern recognition in the grasshopper *Chorthippus biguttulus*: the mechanism of syllable onset and offset detection. *J. Comp. Physiol. A*, 187:255–264, 2001.

O. Barak and M. Tsodyks. Recognition by variance: learning rules for spatiotemporal patterns. *Neural Computation*, 18:2343–2358, 2006.

Horace B. Barlow. Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith, editor, *Sensory Communications*, pages 217–234. MIT Press, Cambridge, MA, 1961.

A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory*, 44:2743–2760, 1998.

D. Bauer. Estimating linear dynamical systems using subspace methods. *Econometric Theory*, 21:181–211, 2005.

J. Benda. *Single Neuron Dynamics - Models Linking Theory and Experiment*. PhD thesis, Humboldt-Universität zu Berlin, 2002.

J. Benda and M. Hennig. Spike-frequency adaptation generates intensity invariance in a primary auditory interneuron. *J. Comp. Neuroscience*, 2007. in print.

J. Benda and A. V. M. Herz. A universal model for spike-frequency adaptation. *Neural Computation*, 15:2523–2564, 2003.

D. Bendor and X. Wang. The neuronal representation of pitch in primate auditory cortex. *Nature*, 436:1161–1165, 2005.

Pietro Berkes and Laurenz Wiskott. Slow feature analysis yields a rich repertoire of complex cells. *Journal of Vision*, 5(6):579–602, 2005.

V. Best, E. Ozmeral, F. J. Gallun, K. Sen, and B. G. Shinn-Cunningham. Spatial unmasking of birdsong in human listeners: Energetic and informational factors. *J. Acoust. Soc. Am.*, 118:3766–3773, 2005.

W. Bialek, I. Nemenman, and N. Tishby. Predictability, complexity and learning. *Neural Computation*, 13(11):2409–2463, 2001.

R. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Inform. Theory*, 18:460–473, 1972.

T. Blaschke, P. Berkes, and L. Wiskott. What is the relation between slow feature analysis and independent component analysis? *Neural Computation*, 18:2495–2508, 2006.

M. Borga. *Learning multidimensional signal processing.* PhD thesis, Linköping Studies in Science and Technology, 1998.

M. Borga and H. Knutsson. A canonical correlation approach to blind source separation. Technical Report LiU-IMT-EX-0062, Department of Biomedical Engineering, Linköping University, 2001.

M. S. Brainard and A. J. Doupe. What songbirds teach us about learning. *Nature*, 417:351–358, 2002.

P. J. Brockwell and R. A. Davis. *Introduction to Time Series and Forecasting.* Springer, 1996.

A. W. Bronkhorst. The cocktail party phenomenon: A review on research on speech intelligibility in multiple-talker conditions. *Acust. Acta Acust.*, 86:117–128, 2000.

J.M. Camhi and M. Hinkle. Attentiveness to sensory stimuli: central control in locusts. *Science*, 175:550–553, 1972.

P. Cariani. Temporal coding of periodicity pitch in the auditory system: an overview. *Neural Plast.*, 6:147–172, 1999.

C. E. Carr and M. Konishi. A circuit for detection of interaural time differences in the brain stem of the barn owl. *J. Neurosci.*, 10:3227–3246, 1990.

A. Cattaneo, L. Maffei, and C. Morrone. Patterns in the discharge of simple and complex visual cortical cells. *Proc. Roy. Soc. B*, 212:279–297, 1981.

M. J. Chacron, B. Lindner, and A. Longtin. Noise shaping by interval correlations increases information transfer. *Phys. Rev. Lett.*, 92:080601, 2004.

G. Chechik, A. Globerson, N. Tishby, and Y. Weiss. Information bottleneck for Gaussian variables. *The Journal of Machine Learning Research*, 6: 165–188, 2005.

C. T. Chen. *Linear System Theory and Design*. Oxford University Press, New York, 1999.

Katrien De Cock and Bart De Moor. Canonical correlations between input and output processes of linear stochastic models, 2002. URL `citeseer.ist.psu.edu/564422.html`.

T.M. Cover and J.A. Thomas. *The elements of information theory*. Plenum Press, New York, 1991.

F. Creutzig and H. Sprekeler. Predictive coding and the slowness principle: an information-theoretic approach. *Neural Comput.*, 20:1026–1041, 2008.

F. Crick. Function of the thalamic reticular complex: the searchlight hypothesis. *Proc. Natl Acad. Sci*, 81:4586–4590, 1984.

S. J. Cruikshank, H. J. Rose, and R. Metherate. Auditory thalamocortical synaptic transmission in vitro. *J Neurophysiol*, 87:361–384, 2002.

I. Csiszar and J. Körner. *Information theory*. Akadémiai Kiadó, Budapest, 1981.

D. Debanne, N. C. Guerineau, and S. M. Gahwiler, B.H.and Thompson. Paired-pulse facilitation and depression at unitary synapses in rat hippocampus: quantal fluctuation affects subsequent release. *J. Physiol.*, 491: 163–176, 1996.

E. deBoer. *Auditory time constants: a paradox? In: Time resolution in auditory systems*, pages 141–158. Berlin: Springer, 1985.

B. C. DeBusk, E. J. DeBruyn, R. K. Snider, J. F. Kabara, and A. B. Bonds. Stimulus-dependent modulation of spike burst length in cat striate cortical cells. *J. Neurophysiol.*, 78:199–213, 1997.

A. G. Dimitrov and J. P. Miller. Neural coding and decoding: Communication channels and decoding. *Network: Computation in Neural Systems*, 12:441–472, 2001.

E. Doi, D. C. Balcan, and M. S. Lewicki. Robust coding over noisy overcomplete channels. *IEEE Transactions on Image Processing*, 16:442–452, 2007.

B. Doiron, A.M.M. Oswald, and L. Maler. Interval coding. II. Dendrite-dependent mechanisms. *J. Neurophysiol.*, 97:2744–2757, 2007.

C. J. Edwards, T. B. Alder, and G. J. Rose. Auditory midbrain neurons that count. *Nat. Neurosci.*, 5:934–936, 2002.

C. J. Edwards, T. B. Alder, and G. J. Rose. Pulse rise time but not duty cycle affects the temporal selectivity of neurons in the anuran midbrain that prefer slow AM rates. *J. Neurophysiol.*, 93:1336–1341, 2005.

J. J. Eggermont and G. M. Smith. Burst-firing sharpens frequency-tuning in primary auditory cortex. *Neuroreport*, 7:753–757, 1996.

M. C. Eguia, M. I. Rabinovich, and H. D. I. Abarbanel. Information transmission and recovery in neural communications channels. *Phys. Rev. E*, 62:7111–7122, 2000.

W. Einhäuser, J. Hipp, J. Eggert, E. Körner, and P. König. Learning viewpoint invariant object representations using a temporal coherence principle. *Biological Cybernetics*, 93(1):79–90, 2005.

M. Elhilali, J. B. Fritz, D. J. Klein, J. Z. Simon, and S. A. Shamma. Dynamics of precise spike timing in primary auditory cortex. *J. Neurosci.*, 24:1159–1172, 2004.

D. Ellis. *Prediction-driven computational auditory scene analysis*. PhD thesis, M.I.T., 1996.

J. Ferbinteanu and M.L. Shapiro. Prospective and retrospective memory coding in the hippocampus. *Neuron*, 40:1227–1239, 2003.

P. Földiak. Learning invariance from transformation sequences. *Neural Computation*, 3:194–200, 1991.

Mathias Franzius, Henning Sprekeler, and Laurenz Wiskott. Slowness leads to place cells. In *Proceedings of CNS 2006*, 2006.

N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby. Multivariate information bottleneck. In *Proceedings of Uncertainty in AI*, 2001.

F. Gabbiani, H. G. Krapp, N. Hatsopoulos, C. Mo, C. Koch, and G. Laurent. Multiplication and stimulus invariance in a looming-sensitive neuron. *J. Physiol.*, 98:19–34, 2004.

F. Gabbiani, W. Metzner, R. Wessel, and C. Koch. From stimulus encoding to feature extraction in weakly electric fish. *Nature*, 384:564–567, 1996.

F. Gabbiani, C. Mo, and G. Laurent. Invariance of angular threshold computation in a wide-field looming-sensitive neuron. *J. Neurosci.*, 21:314–329, 2001.

M Gastpar, B. Rimoldi, and M. Vetterli. To code, or not to code: lossy source-channel communication revisited. *IEEE Trans. Inform. Theory*, 49:1147–1158, 2003.

J. Gautrais and S. Thorpe. Rate coding versus temporal order coding: A theoretical approach. *Biosystems*, 48:57–65, 1998.

I.M. Gelfand and A.M. Yaglom. Calculation of the amount of information about a random function contained in another such function. *Amer. Math. Soc. Trans.*, 12:199–246, 1959.

M. Gevers. A personal view on the development on system identification. *Proc. 13th IFAC Symp. System Identification*, pages 773–784, 2003.

K. Glover. All optimal hankel norm approximations of linear multivariable systems and their $L_\infty$ error bounds. *Int. J. Control*, 39:1115–1193, 1984.

T. Gollisch and A. V. M. Herz. Disentangling sub-millisecond processes within an auditory transduction chain. *PLos Biol. 3:e8*, 2005.

T. Gollisch and M. Meister. Spike latencies in retinal ganglion cells encode spatial image details. In *Cosyne*, 2007.

C. M Gray and D. A. McCormick. Chattering cells: superficial pyramidal neurons contributing to the generation of synchronous oscillations in the visual cortex. *Science*, 274:109–113, 1996.

E. Gray. The fine structure of the insect ear. *Philos. Trans. R. Soc. Lon. B Biol. Sci.*, 243:75–94, 1960.

A. J. Groffen, E. C. Brian, J. J. Dudok, J. Kampmeijer, R. F. Toonen, and M. Verhage. $Ca^{2+}$-induced recruitment of the secretory vesicle protein doc2b to the target membrane. *J. Biol. Chem.*, 279:23740–23747, 2004.

S. Gugercini and A. C. Antoulas. A survey of model reduction by balanced truncation and some new results. *Int.J.Control*, 77:748–766, 2004.

W. Guido, S. M. Lu, J. W. Vaughan, and S. M. Godwin, D. W.and Sherman. Receiver operating characteristics (ROC) analysis of neurons in the cat's lateral geniculate nucleus during tonic and burst response mode. *Vis. Neurosci.*, 12:723–741, 1995.

P. Harremoës and N. Tishby. The information bottleneck revisited or how to choose a good distortion measure. *submitted*, 2007.

R.M. Harris-Warrick and E. Marder. Modulations of neural networks for behavior. *Annu. Rev. Neurosci.*, 14:39–57, 1991.

M.D. Hauser and M. Konishi. *The design of animal communication*. MIT Press, Cambridge, MA, 1999.

R. M. Hecht and N. Tishby. Extraction of relevant speech features using the information bottleneck method. In *Proceedings of InterSpeech*, 2005.

H. Hermansky. Should recognizers have ears? *Speech Communication*, 25: 3–27, 1998.

N. A. Hessler, A. M. Shirke, and R. Malinow. The probability of transmitter release at a mammalian central synapse. *Nature*, 366:569–72, 1993.

B. Ho and R. Kalman. Efficient construction of linear state variable models from input/output functions. *Regelungstechnik*, 14:545–548, 1966.

K. Hoffman. Banach spaces of analytic functions. 1962.

J. J. Hopfield. Transforming neural computations and representing time. *Proc. Natl. Acad. Sci.*, 93:15440–15444, 1996.

J. J. Hopfield and C. Brody. What is a moment? Transient synchrony as a collective mechanism for spatiotemporal integration. *Proc. Natl. Acad. Sci.*, 98:1282–1297, 2001.

J. J. Hopfield, C. Brody, and S. Roweis. Computing with action potentials: toward computation with coupled integrate-and-fire neurons. *Adv. Neural. Inf. Processing*, 10:166–172, 1998.

T. Hosoya, S. A. Baccus, and M. Meister. Dynamic predictive coding by the retina. *Nature*, 436:71–77, 2005.

H. Hotelling. The most predictable criterion. *Journal of Educational Psychology*, 26:139–142, 1935.

D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction, and functional interaction in cat's visual cortex. *J. Physiol.*, 160:106–154, 1962.

D. A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40:1098–1101, 1952.

Jarmo Hurri and Aapo Hyvärinen. Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Computation*, 15(3):663–691, Mar 2003.

M. Ito, H. Tamura, I. Fujita, and K. Tanaka. Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophysiol.*, 73:218–226, 1995.

E. M. Izhikevich. Dynamical systems in neuroscience: The geometry of excitability and bursting. preprint, 2005.

E. M. Izhikevich, N. S. Desai, E. C. Walcott, and F. C. Hoppensteadt. Bursts as a unit of neural information: selective communication via resonance. *Trends Neurosci.*, 26:161–167, 2003.

K. Jacobs, B. Otte, and R. Lakes-Harlan. Tympanal receptor cells of schistocerca gregaria: Correlation of soma positions and dendrite attachment sites, central projections and physiologies. *J. Exp. Zool.*, 283:270–285, 1999.

W. Jacobs. Einige Beobachtungen über Lautäusserungen bei weiblichen Feldheuschrecken. *Z. Tierpsychologie*, 6:141–146, 1944.

L. A. Jeffress. A place theory of sound localization. *J. Comp. Physiol. Psych.*, 41:35–39, 1948.

N.P. Jewell and P. Bloomfield. Canonical correlations of past and future for time series: Definitions and theory. *The Annals of Statistics*, 11:837–847, 1983.

D. Jin. Spiking neural network for recognizing spatiotemporal sequences of spikes. *Phys. Rev. E*, 69:021905, 2004.

R. Johansson and A. Robertsson. On behavioral model identification. *Signal Processing*, 84:1089–1100, 2004.

M. V. Jones and G. L. Westbrook. The impact of receptor desensitization on fast synaptic transmission. *Trends Neurosci.*, 19:96–101, 1996.

B. H. Juang and L. R. Rabiner. Hidden markov models for speech recognition. *Technometrics*, 33:251–272, 1991.

E. R. Kandel, J. H. Schwartz, and T. M. Jessell. *Principles of Neural Sciences*. McGraw-Hill, 2000.

Y. Karklin and M. S. Lewicki. A hierarchical Bayesian model for learning non-linear statistical regularities in non-stationary natural signals. *Neural Computation*, 17(2):397–423, 2005.

T. Katayama. *Subspace methods for system identification*. Springer, 2005.

M. G. Kendall and A. Stuart. *The Advanced Theory of Statistics, volume III*. Griffin, London, 1966.

A. Kepecs, X. J. Wang, and J. Lisman. Bursting neurons signal input slope. *J. Neurosci.*, 22:9053–9062, 2002.

F. A. A. Kingdom, D. R. T. Keeble, and B. Moulden. The perceived orientation of aliased lines. *Vision Research*, 35(19):2759–2766, 1995.

D. H. Klatt. Linguistic uses of segmental duration in english: acoustic and perceptual evidence. *J. Acoust. Soc. Am.*, 59:1208–1221, 1976.

M. Konishi. Birdsong: from behavior to neuron. *Annu. Rev. Neurosci.*, 8: 125–170, 1985.

R. Krahe, E. Budinger, and B. Ronacher. Coding of a sexually dimorphic song feature by auditory interneurons of grasshoppers: the role of leading inhibition. *J. Comp. Physiol A*, 187:977–985, 2002.

R. Krahe and F. Gabbiani. Burst firing in sensory systems. *Nature Neurosci. Rev.*, 5:13–24, 2004.

W.E. Larimore. System identification, reduced order filtering and modeling via canonical variate analysis. *Proc. 1983 American Control Conference*, pages 445–451, 1983.

N. Lemon and R. W. Turner. Conditional spike backpropagation generates burst discharge in a sensory neuron. *J. Neurophysiol.*, 84:1519–1530, 2000.

N.A. Lesica and G.B. Stanley. Encoding of natural scene movies by tonic and burst spikes in the lateral geniculate nucleus. *J. Neurosci.*, 24:10731–10740, 2004.

N.A. Lesica, C. Weng, Jin. J., Yeh C.-I., J.-M. Alonso, and G.B. Stanley. Dynamic encoding of natural luminance sequences by lgn bursts. *Plos Biology*, 4, 2006.

M. S. Lewicki and B. J. Arthur. Hierarchical organization of auditory temporal context sensitivity. *J. Neurosci.*, 16:6987–6998, 1996.

M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12:337–365, 2000.

L.M. Li. Some notes on mutual information between past and future. *J. Time Series Analysis*, 27:309–322, 2005.

L.M. Li and Z. Xie. Model selection and order determination for time series by information between the past and future. *J. Time Series Analysis*, 17: 65–84, 1996.

R.K. Lim, M.Q. Phan, and R.W. Longman. State-space system identification with identified Hankel matrix. Technical Report 3045, Princeton University, 1998.

J. F. Linden, R. C. Liu, M. Sahani, C. E. Schreiner, and M. M. Merzenich. Spectrotemporal structure of receptive fields in areas AI and AAF of mouse auditory cortex. *J. Neurophysiol.*, 90:2660–2675, 2003.

J. E. Lisman. Bursts as a unit of neural information: making unreliable synapses reliable. *Trends Neurosci.*, 20:38–43, 1997.

Y.-H. Liu and X.-J. Wang. Spike-frequency adaptation of a generalized leaky integrate-and-fire model neuron. *J. Comp. Neurosci.*, 10:25–45, 2001.

M. S. Livingstone, D. C. Freeman, and D. H. Hubel. Visual response in V1 of freely viewing monkeys. *Cold Spring Harb. Symp. Quant. Biol.*, 61:27–37, 1996.

L. Ljung. System identification, 1987.

R. Luna, A. Hernandez, C. D. Brody, and R. Romo. Neural codes for perceptual discrimination in primary sensory cortex. *Nat. Neurosci.*, 8:1210–1219, 2005.

C. K. Machens. *Sensory Coding in Natural Environments: Lessons From the Grasshopper Auditory System.* PhD thesis, Humboldt Universität, 2002.

C. K. Machens, H. Schutze, A. Franz, O. Kolesnikova, M. B. Stemmler, B. Ronacher, and A. V. M. Herz. Single auditory neurons rapidly discriminate conspecific communication signals. *Nat. Neurosci.*, 6:341–2, 2003.

C. K. Machens, M. B. Stemmler, P. Prinz, R. Krahe, B. Ronacher, and A. V. M Herz. Representation of acoustic communication signals by insect auditory receptor neurons. *J. Neurosci*, 21:3215–3227, 2001.

D. J. C. Mackay. *Information theory, inference, and learning algorithms.* Cambridge University Press, 2003.

J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Wiley, New York, 1988.

T. Manabe, D. J. Wyllie, D. J. Perkel, and R. A. Nicoll. Modulation of synaptic transmission and long-term potentation: Effects on paired pulse facilitation and epsc variance in the ca1 region of the hippocampus. *J. Neurophysiol.*, 70:1451–1459, 1993.

V. Marquart. *Auditorische Interneurone im thorakalen Nervensystem von Heuschrecken: Morphologie, Physiologie und synaptische Verbindungen.* PhD thesis, Ruhr-Universität Bochum, 1985.

David Marr. *Vision.* W. H. Freeman and Company, 1980.

G. Marsat and G.S. Pollack. A behavioral role for feature detection by sensory bursts. *J. Neurosci.*, 26:10542–10547, 2006.

S. Martinez-Conde, S. L. Macknik, and D. H. Hubel. The function of bursts of spikes during visual fixation in the awake primate lateral geniculate nucleus and primary visual cortex. *Proc. Natl. Acad. Sci.USA*, 99:13920–13925, 2002.

D. A. McCormick and J. R. Huguenard. A model of the electrophysiological properties of thalamocortical relay neurons. *J. Neurophysiol.*, 68:1384–1400, 1992.

W. Metzner, C. Koch, R. Wessel, and F. Gabbiani. Feature extraction by burst-like spike patterns in multiple sensory maps. *J. Neurosci.*, 18:2283–2300, 1998.

B.C. Moore. *An introduction to the psychology of hearing, 4th ed.* Academic Press, New York, 1997.

C. M. Muller and H. Scheich. Contribution of GABAergic inhibition increases the neuronal selectivity to natural sounds in the avian auditory forebrain. *Brain Res.*, 414:376–380, 1987.

R. U. Muller, J. L. Kubie, and J. B Jr. Ranck. Spatial firing patterns of hippocampal complex-spike cells in a fixed environment. *J. Neurosci.*, 7: 1935–1950, 1987.

Jean-Pierre Nadal and Nestor Parga. Redundancy Reduction and Independent Component Analysis: Conditions on Cumulants and Adaptive Approaches. *Neural Computation*, 9(7):1421–1456, 1997. URL `http://neco.mitpress.org/cgi/content/abstract/9/7/1421`.

I. Nelken. Processing of complex stimuli and natural scenes in the auditory cortex. *Curr. Op. Neurobiol.*, 14:474–480, 2004.

I. Nelken, G. Chechik, T. D. Mrsic-Flogel, A. J. King, and J. W. H. Schnupp. Encoding stimulus information by spike numbers and mean response time in primary auditory cortex. *J. Comp. Neurosci.*, 19:199–221, 2005.

I. Nelken, A. Fishbach, L. Las, N. Ulanovsky, and D. Farkas. Primary auditory cortex of cats: feature detection or something else? *Biol. Cybern.*, 89:397–406, 2003.

L. G. Nowak, R. Azouz, M. V. Sanchez-Vives, C. M. Gray, and D. A. McCormick. Electrophysiological classes of cat primary visual cortical neurons in vivo as revealed by quantitative analyses. *J. Neurophysiol.*, pages 1541–1566, 2003.

Goro Obinata and Brian Anderson. *Model Reduction for Control System Design*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2001.

J. Oestreich, N. C. Dembrow, A. A. George, and H. H. Zakon. A "sample-and-hold" pulse-counting integrator as a mechanism for graded memomry underlying sensorimotor adaptation. *Neuron*, 49:577–588, 2006.

B. A. Olshausen and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

M. W. Oram and D. I. Perrett. Time course of neural responses discriminating different views of the face and head. *J. Neurophysiol.*, 68:70–84, 1992.

Randall C. O'Reilly and Mark H. Johnson. Object recognition and sensitive periods: A computational analysis of visual imprinting. *Neural Computation*, 6(3):357–389, 1994.

A.M.M. Oswald, M.J. Chacron, B. Doiron, J. Bastian, and L. Maler. Parallel processing of sensory input by bursts and isolated spikes. *J. Neurosci.*, 24: 402–408, 2004.

A.M.M. Oswald, B. Doiron, and L. Maler. Interval coding. I. Burst interspike intervals as indicators of stimulus intensity. *J. Neurophysiol.*, 97:2731–2743, 2007.

S. Panzeri and A. Treves. Analytical estimates of limited sampling biases in different information measures. *Network*, 7:87–101, 1996.

F. G. Pike, R. M. Meredith, A. W. A. Olding, and O. Paulsen. Postsynaptic bursting is essential for 'hebbian' induction of associative long-term potentation at excitatory synapses in rat hippocampus. *J. Physiol.(Lond.)*, 518: 571–576, 1999.

F. Pongracz, J. D. Poolos, N. P.and Kocsis, and G. M. Shepherd. A model of NMDA receptor-mediated activity in dendrites of hippocampal CA1 pyramidal neurons. *J. Neurophysiol*, 68:2248–2259, 1992.

R. F. Port and J. Dalby. Consonant/vowel ratio as a cue for voicing in english. *Perception & Psychophysics*, 32:141–152, 1982.

R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435: 1102–1107, 2005.

R. P. N. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2:79–87, 1999.

F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek. *Spikes: Exploring the Neural Code*. MIT Press, 1996.

M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neurosci.*, 2:1019–1025, 1999.

J. Rissanen. *Stochastic complexity and statistical inquiry*. World Scientific, 1989.

J. Rissanen. Fisher information and stochastic complexity. *IEEE Trans. Inform. Theory*, 42:40–47, 1996.

A. Rokem, S. Watzl, T. Gollisch, M. Stemmler, A. V. M. Herz, and I. Samengo. Spike-timing precision underlies the coding efficiency of auditory receptor neurons. *J. Neurophysiol.*, 95:2541–2552, 2006.

E. T. Rolls. Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron*, 27:205–218, 2000.

E.T. Rolls and S.M. Stringer. Invariant visual object recognition: a model, with lighting invarance. *J. Phsyiol.*, 100:43–62, 2006.

H Römer. Die Informationsverarbeitung tympanaler Rezeptorelemente von locusta migratoria (Acrididae, Ortophtera). *J. Comp. Physiol.*, A 109: 101–122, 1976.

H. Römer and V. Marquart. Morphology and physiology of auditory interneurons in the metathoracic ganglion of the locust. *J. Comp. Physiol. A*, 155:249–262, 1984.

B Ronacher and R. M. Hennig. Neuronal adaptation improves the recognition of temporal patterns in a grasshopper. *J. Comp. Physiol. A*, 190:311–319, 2004.

D. L. Schacter, D. R. Addis, and R. L. Buckner. Remembering the past to imagine the future: the prospective brain. *Nat. Rev. Neurosci.*, 8:657–661, 2007.

O. Schwartz, T. J. Sejnowski, and P. Dayan. Soft mixer assignment in a hierarchical generative model of natural scene statistics. *Neural Computation*, 18(11):2680–2718, 2006.

K. Sen, F. E. Theunissen, and A. J. Doupe. Feature analysis of natural sounds in the songbird auditory forebrain. *J Neurophysiol*, 86(3):1445–1458, 2001.

M. N. Shadlen and W. T. Newsome. Noise, neural codes and cortical organization. *Curr.Opin.Neurobiol.*, pages 569–579, 1994.

M. N. Shadlen and W. T. Newsome. The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding. *J.Neurosci*, pages 3870–3896, 1998.

C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. Journal*, pages 379–423 and 623–656, 1948.

J. Shaw. *Unifying Perception and Curiosity.* PhD thesis, University of Rochester, 2006.

S.M. Sherman. Tonic and burst firing: dual modes of thalamocortical relay. *Trends Neurosci.*, 24:112–126, 2001.

N. Slonim and N. Tishby. Agglomerative information bottleneck. In *NIPS*, 1999.

S. Soatto and A. Chiuso. Dynamic data factorization. Technical Report UCLA-CSD- 010001, Department of Computer Science, UCLA, 2000.

W. R. Softky and C. Koch. The highly irregular firing of cortical cells is inconsistent with temporal integration of random epsps. *J. Neurosci.*, pages 334–350, 1993.

T. Sokoliuk. *Neuroanatomische Untersuchungen der Hörbahn von* Chorthippus biguttulus. PhD thesis, Friedrich-Alexander-Universität Erlangen Nürnberg, 1992.

T. Sokoliuk, A. Stumpner, and B. Ronacher. GABA-like immunoreactivity suggests an inhibitory function of the thoracic low-frequency neuron (TN1) in Acridid grasshoppers. *Naturwissenschaften*, pages 223–225, 1989.

M. V. Srinivasan, S. B. Laughlin, and A. Dubs. Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 216(1205):427–459, 1982.

R. Stengel. *Optimal control and estimation.* Dover Publications, New York, 1994.

J. Stone and A. Bray. A learning rule for extracting spatio-temporal invariances. *Network: Computation in Neural Systems*, 6(3):429–436, 1995.

M. Stopfer, V. Jayaraman, and G. Laurent. Intensity versus identity coding in an olfactory system. *Neuron*, 39:991–1004, 2003.

A. Stumpner. *Auditorische thorakale Interneurone von* Chorthippus biguttulus *L.: Morphologische und physiologische Charakterisierung und Darstellung ihrer Filtereigenschaften für verhaltensrelevante Lautattrappen.* PhD thesis, Universität Erlangen–Nürnberg, 1988.

A. Stumpner and B. Ronacher. Auditory interneurones in the metathoracic ganglion of the grasshopper *Chorthippus Biguttulus.* I. Morphological and physiological characterization. *J. exp. Biol.*, 158:391–410, 1991.

A. Stumpner and B. Ronacher. Neurophysiological aspects of song pattern recognition and sound localization in grasshoppers. *Amer.Zool.*, 34:696–705, 1994.

A. Stumpner, B. Ronacher, and O. von Helversen. Auditory interneurones in the metathoracic ganglion of the grasshopper *Chorthippus biguttulus.* II. Processing of temporal patterns of the song of the male. *J. exp. Biol.*, 158: 411–430, 1991.

A. Stumpner and D. von Helversen. Evolution and function of auditory systems in insects. *Naturwissenschaften*, 88:159–170, 2001.

H. A. Swadlow and A. G. Gusev. The impact of 'bursting' thalamic impulses at a neocortical synapse. *Nature Neurosci.*, 4:402–408, 2001.

A. Y. Tan, L. I. Zhang, M. M. Merzenich, and C. E. Schreiner. Tone-evoked excitatory and inhibitory synaptic conductances of primary auditory cortex neurons. *J. Neurophysiol.*, 92:630–643, 2004.

A. Thomson. Activity-dependent properties of synaptic transmission at two classes of connections made by rat neocortical pyramidal axons in vitro. *J. Physiol. (Lond.)*, 502:131–147, 1997.

S. Thorpe, A. Delorme, and R. Van Rullen. Spike-based strategies for rapid processing. *Neural Networks*, 14:715–725, 2001.

N. Tishby. The emergence of relevant data representations: an information theoretic approach. In: Lecture notes of the Les Houches summer school 2003. 2005.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proc. of 37th Allerton Conference on communication and computation*, 1999.

M. V. Tsodyks and H. Markram. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proc. Natl. Acad. Sci.USA*, 94:719–723, 1997.

M. V. Tsodyks and H. Markram. Neural networks with dynamic synapses. *Neural Computation*, 10:821–835, 1998.

P. Van Overschee and B. De Moor. N4SID - Subspace algorithms for the identification of combined deterministic - stochastic systems. *Automatica*, 30:75–93, 1994.

Peter van Overschee and Bart De Moor. A unifying theorem for three subspace system identification algorithms. *Automatica*, 31:1853–1864, 1994.

F. Varela. *Principles of biological autonomy.* Elsevier-North Holland, 1976.

G. E. Vates, B. E. Broome, C. V. Mello, and F. Nottebohm. Auditory pathways of caudal telencephalon and their relation to the song system of adult male zebra finches. *J. Comp. Neurol.*, 366:613–642, 1996.

Michel Verhaegen. Identification of the deterministic part of mimo state space models given in innovations form from input-output data. *Automatica*, 30: 61–74, 1994.

N. F. Viemeister and G. H. Wakefield. Temporal integration and multiple looks. *The Journal of the Acoustical Society of America*, 90:858–865, 1991.

P. Vitányi and M. Li. Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Trans. Inform. Theory*, 46:446–464, 2001.

A. Vogel, R. M. Hennig, and B. Ronacher. Increase of neuronal response variability at higher processing levels as revealed by simultaneous recordings. *J. Neurophysio.*, 93:3548–3559, 2005.

D. von Helversen. Gesang des Männchens und Lautschema des Weibchens bei der Feldheuschrecke Chorthippus biguttulus (Orthoptera, Acrididae). *J. comp. Physiol.*, 81:381–422, 1972.

D. von Helversen and O. von Helversen. Recognition of sex in the acoustic communication of the grasshopper *Chorthippus biguttulus* (Orthoptera, Acrididae). *J. Comp. Physiol. A*, 180:373–386, 1997.

D. von Helversen and O. von Helversen. Acoustic pattern recognition in a grasshopper: Processing in the time or frequency domain? *Biol. Cybern.*, 79:467–476, 1998.

O. von Helversen, R. Balakrishnan, and D. von Helversen. Acoustic communication in duetting grasshopper: receiver response variability, male strategies and signal design. *Animal Behavior*, 68:131–144, 2004.

O. von Helversen and D. von Helversen. Forces driving coevolution of song and song recognition in grasshoppers. *In Fortschritte der Zoologie Vol.39: Neural Basis of Behavioural Adaptation, eds. K. Schildberger and N. Elsner*, pages 253–284, 1994. New York, G. Fischer.

I. Waldron. Neural mechanism by which controlling inputs influence motor output in the flying locust. *J. Exp. Biol.*, 47:213–228, 1967.

Guy Wallis and Edmund T. Rolls. Invariant face and object recognition in the visual system. *Progress in Neurobiology*, 51(2):167–194, February 1997.

L. Wang, R. Narayan, G. Grana, M. Shamir, and K. Sen. Cortical discrimination of complex natural stimuli: can single neurons match behavior? *J. Neurosci.*, 27:582–589, 2007.

Y. Wang, H.-F. Guo, T. A. Pologruto, F. Hannan, I. Hakker, K. Svoboda, and Y. Zhong. Stereotyped odor-evoked activity in the mushroom body of drosophila revealed by green flourescent protein-based ca$^{2+}$ imaging. *J. Neurosci.*, 24:6507–6514, 2004.

M. Wehr and A. Zador. Visual adaptation as optimal information transmission. *Neuron*, 47:437–445, 2005.

L. Wiskott and T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002.

Laurenz Wiskott. Slow feature analysis: A theoretical analysis of optimal free responses. *Neural Computation*, 15:2147–2177, 2003.

Laurenz Wiskott. Is slowness a learning principle of visual cortex? In *Proc. Japan-Germany Symposium on Computational Neuroscience, Wako, Saitama, Japan, February 1-4*, page 25. RIKEN Brain Science Institute, 2006.

S. Wohlgemuth, D. Neuhofer, and B. Ronacher. Comparing the neuronal encoding in two not closely related grasshopper species: What differs is different? In *31st Göttingen Neurobiology Conference*, 2007.

S. Wohlgemuth and B. Ronacher. Auditory discrimination of amplitude modulations based on metric distances of spike trains. *J. Neurophysiol.*, 97: 3082–3092, 2007.

R. Wyss, P. König, and P. F. M. Verschure. A model of the ventral visual system based on temporal stability and local memory. *Plos Biology*, 4(5): e120, 2006.

Wei Yu, Wonjong Rhee, Stephen Boyd, and John M. Cioffi. Iterative waterfilling for gaussian vector multiple-access channels. *IEEE Transactions on Information Theory*, 50:145–152, 2004.

K. Zhou, J. Doyle, and K. Glover. *Robust and Optimal Control*. Prentice Hall, 1996.

# Acknowledgement

In my PhD time, I learnt that scientific accomplishment is always not only intellectual but also social in nature and is based on both informal discussions and formal assistance. Most of all, I thank Andreas Herz for his guidance and his sincere support in advancing this thesis and in trying to teach me how to be sufficiently accurate. It was a great pleasure to work with all members of this group, especially with Jan Benda who was always available for discussing general and detailed matters. I am indebted to Susanne Schreiber and Martin Stemmler for critical comments and warm support. Many more people have contributed to providing a great working environment in room 1309 and beyond.

I would like to thank Sandra Wohlgemuth and Bernhard Ronacher for always insisting on taking the whole biological system into account and for sharing their vast knowledge with me. In fact, Sandra Wohlgemuth performed the experiments that form the basis of the first chapters of this thesis.

During my stay in Jerusalem, I profited most from Naftali Tishby who facilitated a great scientific and adventuresome stay in Israel and, crucially, directed my interest into an information-theoretic analysis of dynamical systems. Amir Globerson is responsible for providing me with a first insight on some important mathematical concepts.

Back in Germany, it was ample delight to work together with Henning Sprekeler. I notably acknowledge very supportive discussions with Laurenz Wiskott, always sharp-minded.

Boehringer Ingelheim Fonds and the German National Merit Foundation successively funded my years in Berlin and Jerusalem.

I would like to thank my parents for decades of unconditional support and my dear WG, Veronika Huber, Franka Bindernagel and Stefan Feuerhahn for complementary lifetime outside of academia.

Felix Creutzig Berlin, September 2007

# Selbständigkeitserklärung

Hiermit erkläre ich, dass ich die Dissertation selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe.

Ich habe mich nicht anderwärts um einen Doktorgrad beworben.

Ich bin in Kenntnis der zugrundeliegenden Promotionsordnung.

Berlin, den 3. September 2007